$$V = \int \rho \vec{E} \cdot d\vec{l}$$

$$Z_0 = \sqrt{\frac{L'}{C'}}$$

$$\oint_C \vec{H} \cdot d\vec{l} = \sum I$$

# INTRODUCTORY
# ELECTROMAGNETICS

$$\vec{F} = Q\vec{E}$$

$$\nabla^2 \vec{E} - \omega^2 \epsilon \mu \vec{E} = 0$$

## Zoya Popović · Branko D. Popović

**Michael Faraday** (1791–1867), English chemist and physicist, self-educated from books he was binding to earn a living. An extremely gifted experimentalist, he formulated the law of electromagnetic induction, invented the first dynamo, formulated the fundamental laws of electrolysis, and discovered benzene. Even though he had no formal education, he became the Director of the Royal Institute at age 34 and is certainly one of the greatest scientists ever.

**Joseph Henry** (1791–1878), American professor of philosophy at Princeton, discovered electromagnetic induction independently of Faraday. He invented and operated the first telegraph and discovered self-inductance. He was the first director of the Smithsonian Institution.

**James Clerk Maxwell** (1831–1879), Scottish physicist, the greatest name in classical electromagnetism. He unified the four fundamental laws discovered experimentally by his predecessors by adding the abstract notion of displacement current that enables theoretically wave propagation (described in his famous *Treatise on Electricity and Magnetism*). He predicted theoretically the exact speed of light. He was the first professor of experimental physics at Cambridge. A large portion of his life was dedicated to astronomy, and while investigating Saturn's rings, he formulated the kinetic theory of gases. He was one of the rare scientists who was a brilliant mathematician and experimentalist.

**Heinrich Rudolf Hertz** (1857–1894), German physicist and the first radio and microwave engineer. An ingenious experimentalist as well as theoretician, he demonstrated radio-wave propagation, antennas, microwave sources, polarizers, reflector antennas, first coaxial cable, and many other high-frequency components as they are used today. He discovered the photoelectric effect, for the explanation of which Einstein received the Nobel Prize. He died at the young age of 37 but accomplished more than most long-lived scientists.

**Nikola Tesla** (1856–1943), American inventor, the son of a Serbian priest and a gifted mother who invented many gadgets to help her do housework. A brilliant experimentalist with no complete formal education, he invented the rotating magnetic field, the induction motor (billions of which are used at any moment), and wireless transmission. Tesla designed the first hydroelectric power plant on the Niagara Falls using his three-phase system for ac generation and transmission. He had more than 100 patents, some of which are still under U.S. government secrecy order.

# INTRODUCTORY ELECTROMAGNETICS

Zoya Popović
*University of Colorado*

Branko D. Popović
*University of Belgrade*

To Stefania,
with best wishes for
a long and fruitful
career in EM-related
fields.
I hope you will look
back on our years of
collaboration with fond
memories.
Zoya, 08/2000

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness.

# Contents

## PART 3: SLOWLY TIME-VARYING ELECTROMAGNETIC FIELDS

## 14. Electromagnetic Induction and Faraday's Law          238

## 15. Inductance          265

## 16. Energy and Forces in the Magnetic Field          278

## 17. Some Examples and Applications of Time-Invariant and Slowly Time-Varying Magnetic Fields          299

## 21. Uniform Plane Waves 393

## 22. Reflection and Refraction of Plane Waves 411

## 23. Waveguides and Resonators 432

## 24. Fundamentals of Electromagnetic Wave Radiation and Antennas 457

# Appendix 8:  Theory of Lossless Metallic Waveguides    **539**

# Bibliography    **545**

# Index    **548**

# *Preface*

This text is primarily an intermediate level one-semester textbook in electromagnetic fields, but it can also be used as a two-quarter or two-semester text. Although vector calculus and basic physics are prerequisites, the book is practically self-contained. It is written for engineering and physics students, focusing on physical principles but also applying them to examples from engineering practice.

Below are some points we followed in writing *Introductory Electromagnetics*.

1. The electrical-engineering curricula in most schools are expanding every new academic year. Fundamental subjects, such as electromagnetics, are being covered with fewer hours and in some schools are even being eliminated. Although we believe that this does not benefit future electrical engineers, it is a reality one has to accept. Therefore, we have carefully selected the topics covered in the text to reflect current needs and have stripped it of all less important details.

2. Computers and software tools are now available for solving a large variety of problems. Thus, we feel that it is imperative for future engineers to understand the problems, not so much to be able to perform analytical manipulation of the equations. This textbook stresses the physical basis of applied electromagnetism, including only the necessary minimum of mathematics, which is derived as needed.

3. This text is oriented toward explaining concepts related to what electrical engineers use most frequently—circuit theory. It is our experience that students at the junior level have a better knowledge of circuits than of mathematics and that they need to develop an understanding of where circuit theory comes from. After grasping Kirchoff's and Ohm's laws and understanding distributed capacitance and inductance (Chapters 1 to 17), students learn how circuit theory can be expanded to transmission-line theory, or the wave equation in one dimension (Chapter 18). Subsequently, they learn how this current and voltage-based electromagnetic wave theory can be generalized to waves in three dimensions described by the electric and magnetic field vectors (Chapters 19 to 25).

4. This book includes 25 chapters and 8 appendices. Most of the early chapters are short; they get progressively longer as the knowledge base increases. We believe that short chapters, with clearly marked sections and subsections, make the text clearer and are not intimidating to the reader. In addition, this organization will make it easier for instructors to tailor the lectures to meet their needs.

5. The applications of electromagnetic fields in electrical engineering are becoming progressively more versatile. Many books cover applications of electromagnetic theory; however, in this text, we consider a limited number of applications that are carefully chosen in such a way that they can be understood more than just superficially, which can help the reader solve problems he/she may encounter in the future. The applications are grouped in Chapters 11, 17, and 25, and they combine concepts from all the preceding chapters. We strongly believe that real knowledge is acquired by connecting material studied in different chapters and that practical applications naturally integrate this knowledge, giving it a new depth.

6. We agree with those who state that "examples, questions, and problems make a course." For this reason, we include a large number of examples. At the end of each chapter, questions help the reader to grasp the basic concepts. Carefully selected problems (20 to 40 per chapter) follow the questions.

7. In the authors' opinion, it is extremely useful for students to have a supplement with solved problems, so that they can see what a correct solution should be like. Although students would like to have answers or hints to all questions and problems, it is important that they find solutions themselves. We offer a compromise in *Practice Problems and Labs*, an integral but physically separate part of the text. The supplement provides three answers to questions and three results for problems, one of which is correct. (In some instances, a hint how to approach the problem is given instead.) It also contains short introductory chapter summaries of basic physical theory and equations and units needed to solve all problems in that chapter.

8. Our students have asked us to include some simple laboratories that have been offered as part of this course at the University of Colorado. They told us that, when they did the experiments, the equations that we studied in class came to life. Thus, the supplement includes several very simple and inexpensive experiments. The experiments are designed to use equipment that every electrical engineering and physics department already has: simple oscilloscopes, function

generators, multimeters, and power supplies. If the instructor has the energy, time, and interest to have the students perform some or all of these experiments, the students will benefit. If time or equipment for labs is not available, the instructor might consider using selected experiments as demos, topics for independent study, or just examples on the blackboard. The prelab homework problems can be used as regular problems.

9. This textbook is written by two professors who together have a total of 50 years of teaching experience, both in the United States and in Europe. Both are active in applied electromagnetics research, advise about 20 graduate students, and have many industrial collaborators. The book is also written by a father and a daughter, one contributing experience and the other an outlook to the future.

## Suggested Syllabi and Use of Text

The authors' experience is that the majority of the material contained in this text can be covered in a standard one-semester course (three hours a week, for 15–16 weeks) or in a two-quarter or two-semester course with a greater total number of hours. The instructor can easily decide which parts of the text to skip, or which problems to incorporate into the lectures, to tailor the course for the particular profile of the school and/or students. Suggested below are syllabi for a one-semester, two-quarter (10 weeks each) and two-semester course.

### One-Semester Course Outline

| WEEK | CHAPTERS | WEEK | CHAPTERS | WEEK | CHAPTERS |
|------|----------|------|----------|------|----------|
| 1 | 1–3 | 6 | 13–14 | 11 | 22–23 |
| 2 | 4–5 | 7 | 15–16 | 12 | 23 |
| 3 | 6–7 | 8 | 17–18 | 13 | 24 |
| 4 | 8–10 | 9 | 18–19 | 14 | 25 |
| 5 | 11–13 | 10 | 19–21 | 15 | Review |

### Two-Quarter Course Outline

| WEEK | FIRST-QUARTER CHAPTERS | SECOND-QUARTER CHAPTERS |
|------|------------------------|-------------------------|
| 1 | 1–3 | 17 |
| 2 | 4–5 | 18 |
| 3 | 6–7 | 19–20 |
| 4 | 8–9 | 21 |
| 5 | 10 | 22 |
| 6 | 11–12 | 23 |
| 7 | 13–14 | 24 |
| 8 | 15 | 25 |
| 9 | 16 | Review |
| 10 | Review | Review |

For a two-semester course, the outline is straightforward: complete coverage of Chapters 1–17 in the first semester and Chapters 8–25 in the second, whereby one lecture every other week could be devoted to review or additional examples or problem solving.

*A few additional notes:*

Both at the University of Colorado and at the University of Belgrade, recitation sessions exist in addition to the lectures, and they were found to be extremely useful. If they are not available, a heavier load of homework can replace them. In the authors' opinion, out of the available questions and problems (a total of about 1200), the students should be required to answer at least 200 questions and solve 150 problems chosen by the instructor. This would guarantee a reasonable level of understanding and applicational ability.

It might be interesting for the instructors to know that the authors have used the questions in class competitions, as well as in the first part of every test and exam. They have also incorporated one two-hour lab per week at the University of Colorado, where the prelab problems have to be completed by the students before the beginning of each lab, and the lab report consists of answering about a dozen questions during the lab session. We have thoroughly enjoyed the labs and believe the many students who say that they find them very useful.

## Acknowledgments

The text obtained its final form during the stay of Branko D. Popović at the University of Colorado as a Visiting Professor. He was kindly asked to teach the junior-level electromagnetic fields course (which Zoya Popović teaches often) using rough drafts of some parts of the book. Thus, both authors were able to obtain precious feedback from the students concerning the book's final organization. We are indebted to Professor Renjeng Su, current Chairman of the Department of Electrical and Computer Engineering at the University of Colorado, Boulder, and to Professor Edward Kuester, for their kind effort to enable us to work on the final version of the book at the same physical coordinates. We also thank the faculty in the electromagnetics group at the University of Colorado—Ed Kuester, K. C. Gupta, John Dunn, and Melinda Piket-May—for many useful suggestions and helpful technical discussions. A former student of Branko Popović in Belgrade, Dr. Branislav M. Notaroš, now a faculty member at the University of Massachusetts, Dartmouth, contributed significantly to the solutions in the workbook. The authors would also like to express their gratitude to graduate students at the University of Colorado—Todd Marshall, Manoja Weiss, Michael Forman, Joe Tustin, Shawn Stone, and Jan Peeters-Weem—for being excellent teaching assistants and helping with the development of the labs and to their administrative assistant Helen Frey for being a savior many times. Zoya Popović thanks her husband Professor Dana Anderson for letting her occasionally try out her EM teaching in the physics department and for his love and friendship. Finally, we thank Olya Popović, our mother and wife, respectively, for making us

who we are and our children, sisters, and grandchildren—Sofia, Milica, Nina, and Lena—for helping us keep things in perspective and for their many smiles.

# *Note to the Student*

To quote one of our students, this book can be summarized as "400 [or so] pages on four equations." It is true that Maxwell's equations can describe all the examples, problems, and applications in this book. However, it is also true that these equations were first derived *experimentally* by Coulomb, Ampère, Faraday, and others. Maxwell added one term in Ampère's law that connected all four equations. Therefore, even though electromagnetics might at times seem theoretical to you, please remember its roots. The mathematical apparatus was introduced *to model the physical properties* of electromagnetic fields in a compact way. Whether you are an engineer or a physicist, you will realize in later years the experimental and practical nature of the material covered in this book. We have tried to help you connect the theory to engineering practice by adding chapters on applications, by providing you with a number of practice problems and labs, and by giving realistic values whenever possible.

We would like to give you a few "tips" for learning this material so that you will gain an operational knowledge that will serve you past the final exam, and, we hope, during your entire career. Please treat these "tips" as suggestions: Everyone finds his/her own way of learning a topic. However, our many years of experience and many successful, but also a few unsuccessful, students have provided us with an overview that can benefit you. (We are also certain that your instructor will add other useful suggestions to this list.) So, here are some "recipes for learning electromagnetics."

- Reread each chapter carefully until you can answer most of the questions at the chapter end. If you understand the questions, you can assume you have a good qualitative understanding of the material. You can treat these questions as a game, and study with a friend.

- Make sure you know the basic formulas by heart. This will make it easier for you to use them and understand them, in addition to exercising your brain a little. Write them down on a sheet of paper that you can refer to if you forget them.

- Make sure you know the *units* for all the quantities, as well as some typical values. For example, you should know that capacitance is given in farads (F) and that you probably cannot go to a store and ask for a 2-farad capacitor off the shelf. In addition, units can help you verify many of your solutions, if you know the relationship among the different units (i.e., if you know the basic laws by heart).

- Draw the problems before doing them—many of them are based on physical objects that can be sketched. You will find that, once you know how to sketch a problem, you are half way to solving it.

- Do not get scared by the math. There are a limited number of mathematical tools that you need for this material, and you will master them by the end of the course. It may be a good idea to read through the math appendices first, although some contain more material than you may need.

- As you are studying, try to think of how the material connects with other courses you have taken or are taking now. This will be relatively easy to do for your circuits classes, but you should be able to explain many things in your other courses as well. For example, if you see a "glitch" on your pulse in a digital circuit, it may come from capacitive coupling between two wires or pc-board traces or from the input capacitance of your oscilloscope. Or, you might see a loading effect on a cable that you have not terminated with the right load. You will probably also gain a better understanding of what linear systems mean fundamentally and how time and frequency domain are connected. These are just a few examples of important fundamental concepts that you will learn in this course and that you will use, in one form or another, throughout your career.

- Start studying on time. (You already know this, but it does not hurt to remind you.) This topic might appear easy in the beginning, but *every chapter builds on the previous one*, and so it is important to keep pace. To help you with this, we made the chapters short (except for a few that are technically not easy to divide).

Finally, remember that many people find this topic extremely interesting—but not until they have learned and understood the basics. After you accomplish that, you will have a powerful tool: not only the knowledge of electromagnetics but also a way of thinking that is different from that used in your other classes, as well as techniques that you will be able to apply elsewhere. We hope you work as hard and enjoy yourselves at least as much as we have while preparing this book.

We would like to thank our students from the University of Colorado and the University of Belgrade for putting up with endless lecture notes, correcting many mistakes (true, for extra credit), and being (mostly) enthusiastic about learning. Special thanks to those many students who took time from their busy job schedules after graduating to give us feedback on how they are using this material at work and to thank us for helping them enjoy their professional lives.

*Boulder, Colorado, July 1999*
*Zoya Popović, Associate Professor, University of Colorado*
*Branko D. Popović, Professor, University of Belgrade, Yugoslavia*

"I keep six honest serving-men
(They taught me all I knew);
Their names are What and Why and When
And How and Where and Who."

Rudyard Kipling

# 1

# *Electromagnetics Around Us: Some Basic Concepts*

## 1.1    Introduction

*Electromagnetics* is a brief name for the subject that deals with the theory and applications of electric and magnetic fields. Its implications are of fundamental importance in almost all segments of electrical engineering. Limitations on the speed of modern computers, the range of validity of electrical circuit theory, and the principles of signal transmission by means of optical fibers are just a few examples of topics for which knowledge of electromagnetics is indispensable. Electricity and magnetism also affect practically all aspects of our lives. Probably the most spectacular natural manifestation of electricity is lightning, but without tiny electrical signals buzzing through our nervous system we would not be what we are, and without light (an electromagnetic wave) life on our planet would not be possible.

The purpose of this chapter is to give you a glimpse of what you will learn in this course and how powerful this knowledge is. You will find that you are familiar with some of the information. However, you may also find that some concepts or equations mentioned in this chapter are not easy to understand. Don't let this problem bother you, because we will explain everything in detail later. What is expected at this point is that you refresh some of your knowledge, note some relationships,

1

get a rough understanding of the unity of electricity and magnetism, and above all, understand how important this subject is in most of electrical engineering.

Electromagnetic devices are almost everywhere: in TV receivers, car ignition systems, elevators, and mobile phones, for instance. Although it may sometimes be hard to see the fundamental electromagnetic concepts on which their operation is based, you certainly cannot design these devices and understand how they work if you do not know basic electromagnetic principles.

In this chapter we first look at a few examples that show how the knowledge you will gain through this course can help you understand, analyze, and design different electrical devices. We will start with a typical office, which is likely to have a computer and a printer or a copier. We will list the different components and mechanisms inside the computer, relating them to chapters we will study later in the course. You may not yet understand what all the words mean, but that should not alarm you. During the course we will come back to these examples, each time with more understanding.

*Questions and problems:*    Q1.1 to Q1.3

## 1.2    Electromagnetics in Your Office

Let us consider a personal desktop computer connected to a printing device and list the different components and mechanisms that involve knowledge of electricity or magnetism (Fig. 1.1).

1. The computer needs energy. It has to be plugged into a wall socket—that is, to an ac voltage generator. An ac voltage generator converts some form of energy into electrical energy. For example, hydroelectric power plants have large
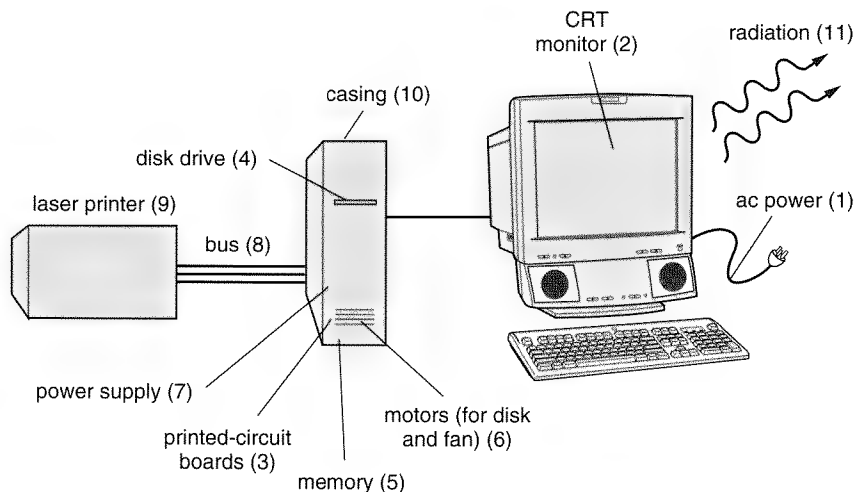


Figure 1.1 A personal desktop computer plugged into the wall socket and connected to a printer

generators in which the turbines, powered by water, produce rotating magnetic fields. We will study in Chapter 14 how such a generator can be built. These generators are made of copper conductors and iron or other magnetic materials, the properties of which we will study in Chapter 13.

2. Most desktop computers use a cathode-ray tube (CRT) monitor. In Chapter 17, we will explain how a CRT works. It involves understanding charge motion in electric and magnetic fields. Basically, a stream of electrons (negatively charged particles) is accelerated by an electric field and then deflected by a magnetic field, to trace a point on the front surface of the monitor and, point by point, a full image. The CRT runs off very high voltages, so the 110-V (or 220-V) socket voltage needs to be transformed into a voltage of a few kilovolts, which accelerates the electron beam. This is done using a magnetic circuit, or transformer, which we will study in Chapters 13 and 17.

3. The computer cabinet, or system unit, contains numerous printed-circuit boards. They contain conductive traces (Chapter 6) on dielectric substrates (Chapter 7); chips with many transistors, which are essentially charge-control devices (Chapter 7); and elements such as capacitors, resistors, and inductors (Chapters 8, 10, and 15). Signals flowing through the board traces couple to each other by electric (capacitive) and magnetic (inductive) coupling, which we will study in Chapters 8 and 15.

4. Many disks are read by magnetic heads from ferromagnetic traces. This is the topic of Chapters 14 and 17.

5. Computer memory used to be magnetic, built of small ferromagnetic toruses (Chapter 17). Now it is made of transistors, which serve as charge storage devices. We describe this mechanism in Chapter 8.

6. Inside the computer a motor operates the cooling fan. A motor converts electric energy to mechanical energy.

7. The semiconductor chips in the computer need typically 5 V or 3 V dc, instead of the 60-Hz 110 V (or 50-Hz 220 V) available from the socket. The power supply inside the computer performs the conversion. It uses components such as inductors, capacitors, and transformers, which we have already listed above.

8. The computer is connected to the printer by a multi-wire bus. The different lines of the bus can couple to each other capacitively (Chapter 8) and inductively (Chapter 15), and the bus can have an electromagnetic wave traveling along it, which we will discuss in Chapters 18, 23, and 25.

9. The printer will probably be a laser printer or an ink-jet printer. The laser printer operates essentially the same way as a copier machine, which is based on recording an electrostatic charge image and then transferring it to paper. The ink-jet printer is also an electrostatic device, and we will describe operations of both types of printers in Chapter 11.

10. The computer parts are shielded from outside interference by their metal casings. We are all bathing constantly in electromagnetic fields of different frequencies and intensities, which have different penetration properties into different materials (Chapter 20). However, some of the computer parts sometimes act

as receiving antennas (Chapter 24), which couple the interference onto signal lines, causing errors. This is called *electromagnetic interference (EMI)*. The regulations that are imposed on frequency band allocations, allowed power levels, and shielding properties are generally referred to as *electromagnetic compatibility (EMC)* regulations.

11. The computer also radiates a small amount of energy—that is, it acts as a transmitting antenna at some frequencies. We will study basic antenna principles in Chapter 24.

12. Finally, when we use the computer we are (we hope) thinking, which makes tiny voltage impulses in our neurons. Since our cells are mostly salty water, which is a liquid conductor, the current in the neurons will roughly have the same properties as the one through wire conductors.

## 1.3    Electromagnetics in Your Home

Now let us look at some uses of electromagnetics in your home. We know that most household appliances need ac voltage for their operation and that most of them (for example, blenders, washers, dryers, fans) contain some kind of electric motor. Both motors and generators operate according to principles that are covered in the third part of this book. An electric oven, as well as any other electric heating element (such as the one in a hair dryer or curling iron), operates according to Joule's law, which is covered in Chapter 10. Your washer, dryer, and car have been painted using electrostatic coating techniques, which we will briefly describe in Chapter 11.

Your TV receiver contains a cathode-ray tube, which, as we mentioned earlier, is described in Chapter 17. It is connected to the cable distribution box with a coaxial cable, a transmission line we will study throughout this book (Chapter 18). A transmission line supports an electromagnetic wave (Chapters 21 and 22). A similar wave traveling in free space is captured by an antenna, which you might also own. It could be a simple "rabbit ears" wire antenna or a highly directional reflector (dish) antenna. Basic antenna principles are covered in Chapter 24. Your cordless phone also contains an antenna, as well as high-frequency (rf) circuitry. All these applied electromagnetics topics are discussed in higher level courses in this field. Some of these applications are briefly described in Chapter 25 in the context of communications engineering.

A microwave oven is essentially a resonant cavity (Chapter 23), in which electromagnetic fields of a very high frequency are contained. The energy of these fields (Chapter 19) is used to heat up water (Chapter 25), whose molecule has a rotational resonance in a broad range around the designated heating frequency of 2.45 GHz. Thus the energy of the electromagnetic wave is transformed into kinetic energy of the water molecules, which on average determines the temperature of water. Because a large percentage of most foods is water, this in turn determines the food temperature.

Many other examples of electromagnetic phenomena occur in everyday life— light, which enables you to read these pages, is an electromagnetic wave. White light covers a relatively narrow range of frequencies, and our eyes are frequency-dependent sensors of electromagnetic radiation (that is, antennas for the visible part of the electromagnetic spectrum).

## 1.4    A Brief Historical Introduction

A tour through the historical development of the knowledge of electricity and magnetism reveals that this seemingly theoretical subject is entirely based on experimentally discovered laws of nature.

### 1.4.1   THE BEGINNING

When and where were the phenomena of electricity and magnetism first noticed? Around 600 B.C., the Greek philosopher and mathematician Thales of Miletus found that when amber was rubbed with a woolen cloth, it attracted light objects, such as feathers. He could not explain the result but thought the experiment was worth writing down. Miletus was at the time an important Greek port and cultural center. Ruins of Miletus still exist in today's Turkey, shown on the map in Fig. 1.2. Some 20 km from Miletus is an archaeological site called Magnesia, where the ancient Greeks first found magnetite, a magnetic ore. They noticed that lumps of this ore attracted one another and also attracted small iron objects. The word *magnet* comes from the name of the place where this ore was found.
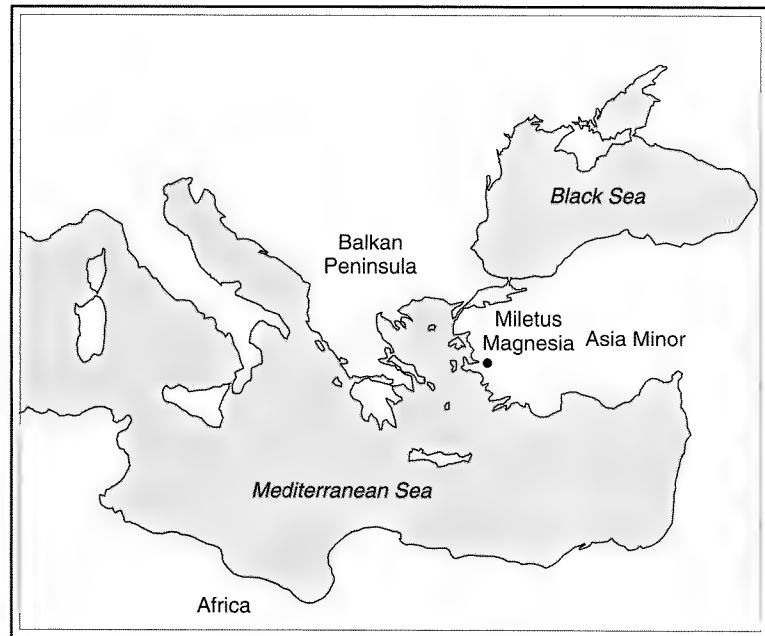


Figure 1.2 Map of the Mediterranean coast. Until Roman times, most coast colonies were Greek. Miletus was an important port and cultural center, connected by a 16-km marble road, lined with statues, to the largest Greek temple ever built (but never finished), at Didime.

Thus the first manifestations of both electricity and magnetism were noticed by the ancient Greeks at about the same time and at almost the same place. This coincidence was in a way an omen: we now know that electricity and magnetism are two facets of the same physical phenomenon.

### 1.4.2   CHRISTENING OF ELECTRICITY 22 CENTURIES LATER

There is no evidence that people thought about what Thales had observed for the next 2200 years. Around the year 1600 a physician to Queen Elizabeth I, William Gilbert, repeated Thales's experiments in a systematic way. He christened "electricity" from the Greek word for amber, *electron*, in honor of Thales's experiments. He rubbed different materials with woolen or silk cloth and concluded that some repel each other, and others are attracted after they are rubbed. We now know that when a piece of amber is rubbed with wool, some electrons (negative charges) from the wool molecules hop over to the amber molecules and therefore the amber has extra electrons. We say that the amber is *negatively charged*. The wool has fewer electrons, which makes it also different from neutral, and we say it is *positively charged.*

### 1.4.3   POSITIVE AND NEGATIVE CHARGES

The terms *positive* and *negative electric charges* were introduced by Benjamin Franklin (around 1750) for no particular reason; he could also have called them red and blue. It turned out, however, that for mathematically describing electrical phenomena, associating "+" and "−" signs with the two kinds of electricity was extremely convenient. For example, electrically neutral bodies are known to contain very large but equal amounts of positive and negative electric charges; the "+" and "−" convention allows us to describe them as having zero total charge.

Why were electrical phenomena not noticed earlier? The gravitational force has been known and used ever since the ancient man poured, for example, water in his primitive container. This time lag can be easily understood if we compare the magnitudes of electrical forces and some other forces acting around us.

### 1.4.4   COULOMB'S LAW

Electrical forces were first investigated systematically by Charles de Coulomb in 1784. By that time it was well established that like charges repel and opposite charges attract each other, but it was not known how this force could be calculated. Using a modified, extremely sensitive torsion balance (with a fine silk thread replacing the torsion spring), Coulomb found experimentally that the intensity of the force between two "point" charges (charged bodies that are small compared to the distance between them) is proportional to the product of their charges ($Q_1$ and $Q_2$ in Fig. 1.3), and inversely proportional to the square of the distance $r$ between them:

$$F_e = k_e \frac{Q_1 Q_2}{r^2}. \tag{1.1}$$
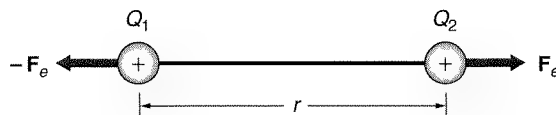
**Figure 1.3** Coulomb's electric force between two particles with charges of the same sign, which are small in size compared to the distance $r$ between them

This is *Coulomb's law*. The unit for charge we use is called a *coulomb* (C). With the distance $r$ in meters and force $F$ in newtons (N), the constant $k_e$ is found to be very nearly $9 \times 10^9$ N·m$^2$/C$^2$. This force is attractive for different charges (one positive and the other negative), and repulsive for like charges (both negative or both positive). The charge of an electron turns out to be approximately $-1.6 \times 10^{-19}$ C.

How large is this force? Let us first look at the formula. If we replace the constant $k_e$ with the gravitational constant $\gamma = 6.67 \times 10^{-11}$ N · m$^2$/kg$^2$, and the charges by the masses, $m_1$ and $m_2$, of the two particles (in kg), the formula becomes that for the gravitational force between the two particles due to their masses:

$$F_g = \gamma \frac{m_1 m_2}{r^2}. \tag{1.2}$$

Let us calculate how the electric force in a hydrogen atom (which has one electron and one proton) compares to the gravitational force. Using the preceding formulas and the data for the masses of an electron and a proton given in Appendix 3, we find that the ratio of the electric to gravitational forces between the electron and the proton of a hydrogen atom is astonishing:

$$\frac{F_e}{F_g} \simeq 10^{39}.$$

We know that atoms of matter are composed of elemental charges that include protons and electrons. If this is the ratio of electric to gravitational force acting between one proton and one electron, we should also expect enormous electric forces acting around us. Yet we can hardly notice them. They include such minor effects as our hair rising after we pull off a sweater. There are simply no appreciably larger electric forces in everyday life. How is this possible? To understand it, let us do a simple calculation.

Assume two students are sitting 1 m apart and their heads are charged. Let us find the force between the two heads, assuming they are point charges (for most students, of course, this is not at all true, but we are doing only an approximate calculation). Our bodies consist mostly of water, and each water molecule has 10 electrons and the same number of protons in one oxygen atom and two hydrogen atoms. Thus we are nothing but a vast ensemble of electric charges. In normal circumstances, the amount of positive and negative charges in the body is practically balanced, i.e., the net charge of which our body is composed is very nearly zero.

### 1.4.5   PERCENTAGE OF EXCESS CHARGE
###         ON CHARGED BODIES

Let us assume, however, that a small percentage of the total electron charge, say 0.1%, exists in excess of the total positive charge. If each head has a volume of roughly $10^{-3}\,\mathrm{m}^3$, and solids and liquids have about $10^{28}\,\mathrm{atoms/m}^3$, each head has on the order of $10^{25}$ atoms. Assume an average of 10 electrons per atom (human tissue consists of various atoms). One tenth of a percent of this is roughly $10^{23}$ electrons/head. Since every electron has a charge of $-1.6 \times 10^{-19}\,\mathrm{C}$, this is an extra charge of about $-1.6 \times 10^4\,\mathrm{C}$. When we substitute this value into Coulomb's law, we find that the force between the two students' heads 1 m apart is on the order of $2 \times 10^{20}$ newtons (N).

How large is this force? The "weight" of the earth, if such a thing could be defined, would be on the order of $10^{20}\,\mathrm{N}$, that is, of same order of magnitude as the previously estimated force between the two students. How is it then possible that we do not notice the electric force? Where did our calculation go wrong? The answer is obvious: we assumed too high a percentage (0.1%) of excess electrons. Since we do not notice electric forces in common life, this tells us that the charges in our world are *extremely well balanced*, i.e., that only a very small percentage of protons or electrons in a body is in excess over the other.

### 1.4.6   CAPACITORS AND ELECTRIC CURRENT

We know that extra charge can be produced by rubbing one material against another. This charge can stay on the material for some time, but it is very difficult to collect from there and put somewhere else. It is of extreme practical importance to have a device analogous to a water container in which it is possible to store charge. Devices that are able to act as charge containers are called *capacitors*. They consist of two conducting pieces known as *capacitor electrodes* that are charged with *charges of equal magnitude but opposite signs*. An example is in Fig. 1.4a.

If the medium between the two electrodes is air, and if many small charged particles are placed there, the electric forces due to *both* electrodes will move the charges systematically toward the electrode of the opposite sign. Such an ordered motion of a large number of electric charges is called the *electric current* because it resembles
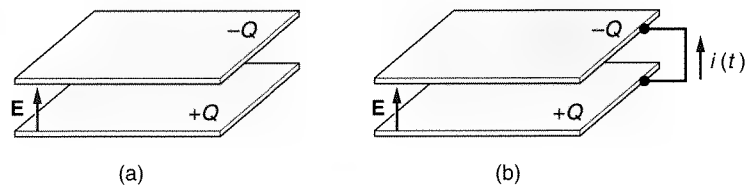


Figure 1.4 (a) A simple capacitor consists of two oppositely charged bodies. (b) If the two capacitor electrodes are connected by a wire, a short flow of charges occurs until the capacitor is discharged.

the current of a fluid. We can get the same effect more easily if we connect the two electrodes by a metallic (conducting) wire. A short flow of electrons in the metal wire will result, until the capacitor is discharged (Fig. 1.4b), i.e., until all of the negative charges neutralize the positive ones. Thus a charged capacitor cannot sustain a permanent electric current.

### 1.4.7 ELECTRIC GENERATORS

This flow of charges, more precisely an effect of this flow, was first noticed around 1790 by Luigi Galvani when he placed metal tweezers on a frog's leg and noticed that the leg twitched. Soon after that, between 1800 and 1810, Alessandro Volta made the first battery—a device that was able to maintain a continuous charge flow for a reasonable time.

   A sketch of Volta's battery is shown in Fig. 1.5. The battery consisted of zinc and copper disks separated by leather soaked in vinegar. The chemical reactions between the vinegar and the two types of metal result in opposite charges on copper and zinc disks. These charges exert a force on freely movable electrons in a wire connecting them, resulting in electric current in the wire. Obviously, the larger these charges, the stronger the force on electrons in the wire. A quantity that is directly proportional to the charge on one of the disks is known as *voltage*. The unit of voltage is the *volt* (V), in honor of Volta. Volta "measured" the voltage by placing two pieces of wire on his tongue (the voltage is about 1 V per cell).

   The chemical reaction that governs the process in a zinc-copper battery that uses a solution of sulfuric acid ($H_2SO_4$) is given by the following equation, assuming
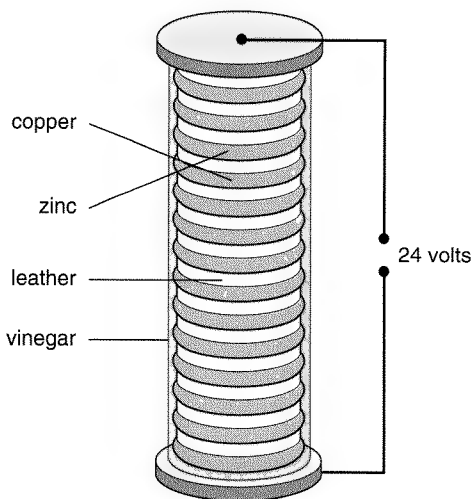


Figure 1.5 Volta's first battery consisted of 24 pairs of copper and zinc disks separated by leather soaked in vinegar.

the end copper (Cu) and zinc (Zn) plates to be connected with a conducting wire:

$$Cu + Zn + 3H_2SO_4 = Zn^{++} + 2SO_4^{--} + Cu^{++} + SO_2 + H_2 + 2H_2O. \qquad (1.3)$$
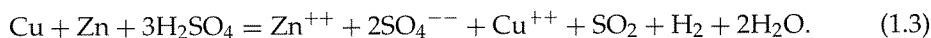
Hydrogen gas molecules ($H_2$) are given off at the copper plate, which loses electrons to the solution and becomes positively charged. Zinc dissolves from the zinc plate, leaving electrons behind. The electrons move through the wire from the zinc to the copper plate, making an electric current. The process stops when the zinc plate is eaten away, or when no more acid is left.

Volta's battery is just one type of *electric generator*. Other chemical generators operate like Volta's battery but with different substances. However, generators can separate positive and negative electric charges, that is, can produce a voltage between their terminals, in many different ways: by a wire moving in a magnetic field; by light charging two electrodes of a specific semiconductor device; by heating one connection of two wires made of different materials; and even by moving charges mechanically (which, however, is extremely inefficient). All electric generators have one common property: they use some other kind of energy (chemical, mechanical, thermal, solar) to separate electric charges and to obtain two charged electrodes.

### 1.4.8  JOULE'S LOSSES

When there is an electric current in a substance, the electric force accelerates charged particles that can move inside the substance (e.g., electrons in metals). After a very short trip, however, these particles collide with atoms within the substance and lose some energy they acquired by acceleration. This lost energy is transformed into heat—more vigorous vibrations of atoms inside the substance. This heat is known as *Joule's heat* or *Joule's losses*.

### 1.4.9  MAGNETISM

As mentioned, the phenomenon of magnetism was first noticed at about the same time as that of electricity. The magnetic needle (a small magnet suspended to rotate freely about a vertical axis) was observed by the Chinese about 120 B.C. The magnetic force was even more mysterious than the electric force. Every magnet *always has two "poles"* that cannot be separated by cutting a magnet in half. In addition, one pole of the magnetic needle, known as its *north pole*, always turns itself toward the north.

People could not understand why this happened. An "explanation" that lasted for many centuries (until about A.D. 1600) was that the north pole of the needle was attracted by the North Star. This does not show, of course, that our ancestors were illogical, for without the knowledge we have today we would probably accept the same explanation. Instead it shows at least two things typical of the development of human knowledge: we like simple explanations, and we tend to take explanations for granted. Whereas the desire to find a simpler explanation presents a great positive challenge, the tendency to take explanations for granted presents a great danger.

The magnetic forces were also studied experimentally by Coulomb. Using long magnets and his torsion balance, he concluded that the magnetic poles exert forces on each other and that these forces are of the same form as those between two point

charges. This is known as the Coulomb force for magnetic poles, and it represents another approach we frequently use in trying to understand things: the use of analogies. We will see shortly that magnetic poles actually do not exist. This example, therefore, demonstrates that we should be careful about analogies and be critical of them.

### 1.4.10  ELECTROMAGNETISM AND ORIGIN OF MAGNETISM IN PERMANENT MAGNETS

Because of Coulomb's law for magnetic poles, magnetism was for some time considered to be separate from electricity but to have very similar laws. Around 1820, however, the Danish physicist Hans Christian Oersted noticed that a magnetic needle is deflected from its normal orientation (north-south) if placed close to a wire with electric current. Knowing that two magnets act on each other, he concluded that a wire with electric current is a kind of magnet, i.e., that *magnetism is due to moving electric charges*. This "magnet" is, of course, different from a piece of magnetic ore (a permanent magnet) because it can be turned on and off and its value can be controlled. It is called an *electromagnet* and has many uses, for example cranes and starter motors.

Soon after Oersted's discovery, the French physicist André Marie Ampère offered an explanation of the origin of magnetism in permanent magnets. He argued that inside a permanent magnet there must be a large number of tiny loops of electric current. He also proposed a mathematical expression describing the force between two short segments of wire with current in them. We will see in a later chapter that this expression is more complicated than Coulomb's law. However, for the particular case of two parallel short wire segments $l_1$ and $l_2$ with currents $I_1$ and $I_2$, shown in Fig. 1.6, *and only in that case*, this expression is simple:

$$F_m = k_m \frac{(I_1 l_1)(I_2 l_2)}{r^2},$$
(1.4)

where $k_m$ is a constant. The direction of the force in the case in Fig. 1.6 (parallel elements with current in the same direction) is *attractive*. It is repulsive if the currents in the elements are in opposite directions. Note that an analogy with electric forces might tempt us to anticipate (erroneously) different force directions than the actual ones.
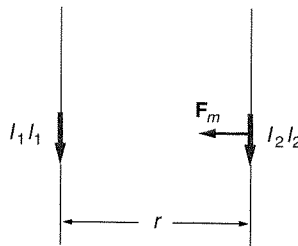


Figure 1.6 Magnetic force between two parallel current elements
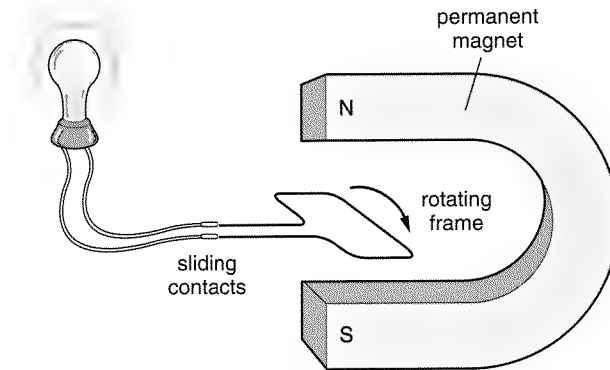
Figure 1.7 A simple generator can be made by turning a wire frame in a dc magnetic field.

### 1.4.11   ELECTROMAGNETIC INDUCTION

The final important physical fact of electricity and magnetism we mention was discovered in 1831 by the British physicist Michael Faraday. He performed experiments to check whether Oersted's experiment was reciprocal, i.e., whether current will be produced in a wire loop placed near a magnet. He did not find that, but he realized that *a current in the loop was obtained while the magnet was being moved toward or away from it.* The law that enables this current to be calculated is known as *Faraday's law of electromagnetic induction.*

As an example, consider a simple generator based on electromagnetic induction. It consists of a wire frame rotating in a time-constant magnetic field, as in Fig. 1.7, with the ends of the frame connected to the "outer world" by means of sliding contacts. Let the sliding contacts be connected by a separate and stationary wire, so that a closed conducting loop is obtained. When the wire frame turns, its position with respect to the magnet varies periodically in time, which induces a varying current in the frame and the wire that completes the closed conducting loop.

*Questions and problems:*    Q1.4 to Q1.16, P1.1 to P1.4, P1.10

## 1.5    The Concept of Electric and Magnetic Field

Let us now assume that we know the position of the charge $Q_1$ in Coulomb's law, but that there are several charges close to charge $Q_1$, of unknown magnitudes and signs and at unknown locations (Fig. 1.8). We cannot then calculate the force on $Q_1$ using Coulomb's law, but from Coulomb's law, and knowing that mechanical forces add as vectors, we anticipate that there will be a force on $Q_1$ proportional to $Q_1$ itself:

$$\mathbf{F}_e = Q_1\mathbf{E}. \tag{1.5}$$
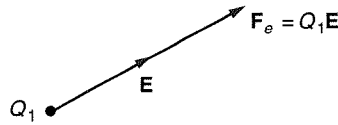
Figure 1.8 The electric field vector, **E**, is defined by the force acting on a charged particle.

(It is customary in printed text to use boldface fonts for vectors, e.g., **r**. In handwriting, vectors are denoted by an arrow above the letter, e.g., $\vec{r}$. A brief survey of vectors is given in Appendix 1.)

This is the definition of the *electric field strength*, **E**. It is a vector, equal to the force on a small charged body at a point in space, divided by the charge of the body. Note that **E** generally differs from one point to another, and that it frequently varies in time (for example, if we move the charges producing **E**). The domain of space where there is a force on a charged body is called the *electric field*. Thus, we can describe the electric field by **E**, a *vector* function of space coordinates (and possibly of time). For example, in a Cartesian coordinate system we would write: $\mathbf{E}(x, y, z, t) = \mathbf{E}_x(x, y, z, t) + \mathbf{E}_y(x, y, z, t) + \mathbf{E}_z(x, y, z, t)$. Obviously, sources of the electric field are electric charges and currents. If sources producing the field are not moving, the field can be calculated from Coulomb's law. This kind of field is termed the *electrostatic field*, meaning "the field produced by electric charges that are not moving."

Consider now Eq. (1.4) for the magnetic force between two current elements and assume that several current elements of unknown intensities, directions, and positions are close to current element $I_1 l_1$. The resulting magnetic force will be proportional to $I_1 l_1$. We know that current elements are nothing but small domains with moving charges. Let the velocity of charges in the current element $I_1 l_1$ be **v**, and the charge of individual charge carriers in the current element be $Q$. The force on the current element is the result of forces on individual moving charge carriers, so that the force on a single charge carrier should be expected to be proportional to $Q\mathbf{v}$. Experimentally, the expression for this force is found to be of the form

$$\mathbf{F}_m = Q\mathbf{v} \times \mathbf{B}, \tag{1.6}$$

where the sign "×" implies the vector, or cross, product of two vectors (Appendix 1). The vector **B** is known as the *magnetic induction vector* or the *magnetic flux density vector*. If in a region of space a force of the form in Eq. (1.6) exists on a moving charge, we say that in that region there is a *magnetic field*.

*Questions and problems:*    Q1.17 to Q1.20, P1.5 to P1.9

## 1.6    The Electromagnetic Field

Faraday's law shows that a time-varying magnetic field produces a time-varying electric field. Is the converse also true? About 1860 the British physicist James Clerk Maxwell stated that this must be so, and he formulated general differential equations

of the electric and magnetic fields that take this assumption into account. Because the electric and magnetic fields in these equations are interrelated in such a manner that if they are variable in time one cannot exist without the other, this resulting field is known as the *electromagnetic field*. These famous equations, known as *Maxwell's equations*, have proven to be exact in all cases of electromagnetic fields considered since his time. In particular, Maxwell theoretically showed from his equations that an electromagnetic field can detach itself from its sources and propagate through space as a field package, known as an *electromagnetic wave*. He also found theoretically that the speed of this wave in air is the same as the speed of light measured earlier by several scientists (for example, Roemer in 1675 estimated it to be about $2.2 \times 10^8$ m/s, and Fizeau in 1849 and Foucault in 1850 determined it to be about $3 \times 10^8$ m/s). This led him to the conclusion that light must be an electromagnetic wave and he formulated his famous electromagnetic theory of light. Maxwell's equations break down, however, at the atomic level because the field quantities used in the equations are averaged over many atoms. Such quantities are called *macroscopic*. (The science that deals with electromagnetic phenomena at the atomic and subatomic levels is called *quantum physics*.)

The first person who experimentally verified Maxwell's theory was the German physicist Heinrich Hertz. Between 1887 and 1891 he performed a large number of ingenious experiments at frequencies between 50 MHz and 5 GHz. At that time, these were incredibly high frequencies. One of his experiments proved the existence of electromagnetic waves. A device that launches or captures electromagnetic waves is called an *antenna*. Hertz used a high voltage spark (intense current in air of short duration, and therefore rich in high frequencies) to excite an antenna at about 60 MHz (Fig. 1.9). This was his transmitter. The receiver was an adjustable loop of wire with another spark gap. When he adjusted the resonance of the receiving antenna to that of the transmitting one, he was able to notice a weak spark in the gap of the receiving antenna. Hertz thus demonstrated for the first time that Maxwell's predictions about
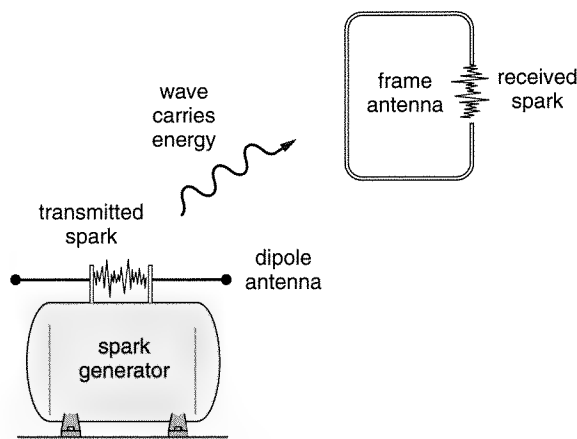


Figure 1.9 Hertz's first demonstration of an electromagnetic wave

the existence of electromagnetic waves were correct. Hertz also introduced the first reflector antennas, predicted the finite velocity of waves in coaxial transmission lines and the existence of standing electromagnetic waves, as well as a number of radio techniques used today. He was, in fact, the first radio engineer.

Electric, magnetic, or electromagnetic fields are present in any device we use in electrical engineering. Therefore, Maxwell's equations should strictly be used for the analysis and design of all such devices. This would be quite a complicated process, however. Fortunately, in many cases approximations that simplify the analysis process are possible. For example, circuit theory is essentially a very powerful and simple approximation of the exact field theory. In the next chapter we look at the interconnection between fields and circuits, and explore briefly the electromagnetic foundations of circuit theory and its limitations.

*Questions and problems:*   Q1.21, Q1.22

## 1.7    Chapter Summary

1. The principal developments in the history of the science of electricity and magnetism began with the ancient Greeks. Key concepts, however, have been described only in the past 400 years.

2. The objects in the world around us are composed of very nearly equal numbers of elemental positive and negative electric charges. The excess charge of one kind over the other can be only an extremely small fraction of the total charge of that kind.

3. Between stationary bodies with excess charges, which we term *charged bodies,* there is a force known as the *electric force.*

4. If there is a force on a charge $Q$ in a region of space of the form $\mathbf{F}_e = Q\mathbf{E}$, we say that an *electric field* exists in that region. The vector $\mathbf{E}$ is known as the *electric field vector.*

5. If charges are moving, there is an additional force acting between them. It is called the *magnetic force.*

6. If there is a force on an electric charge $Q$ moving with a velocity $\mathbf{v}$ in a region of space, of the form $\mathbf{F}_m = Q\mathbf{v} \times \mathbf{B}$, we say that a *magnetic field* exists in that region. The vector $\mathbf{B}$ is known as the *magnetic induction vector* or *magnetic flux density vector.*

7. An electric field that varies in time is always accompanied by a magnetic field that varies in time, and vice versa. This combined field is known as the *electromagnetic field.*

8. The equations that mathematically describe any electric, magnetic, and electromagnetic field are known as *Maxwell's equations.* They are mostly based on experimentally obtained physical laws.

# QUESTIONS

**Q1.1.** What is electromagnetics?

**Q1.2.** Think of a few examples of animals that use electricity or electromagnetic waves. What about a bat?

**Q1.3.** The basis of plant life is photosynthesis, i.e., synthesis (production) of life-sustaining substances by means of light. Is an electromagnetic phenomenon included?

**Q1.4.** What is the origin of the word *electricity*?

**Q1.5.** What is the origin of the word *magnetism*?

**Q1.6.** When did Thales of Miletus and William Gilbert make their discoveries?

**Q1.7.** Why is it convenient to associate plus and minus signs with the two kinds of electric charges?

**Q1.8.** When did Coulomb perform his experiments with electric forces?

**Q1.9.** What is the definition of a capacitor?

**Q1.10.** What is electric current?

**Q1.11.** What are electric generators?

**Q1.12.** What common property do all electric generators have?

**Q1.13.** Describe in your own words the origin of Joule's losses.

**Q1.14.** What is the fundamental cause of magnetism?

**Q1.15.** What is an electromagnet?

**Q1.16.** What did Faraday notice in 1831 when he moved a magnet around a closed wire loop? What did he expect to see?

**Q1.17.** Explain the concept of the electric field.

**Q1.18.** Define the electric field strength vector.

**Q1.19.** Explain the concept of the magnetic field.

**Q1.20.** Define the magnetic induction (magnetic flux density) vector.

**Q1.21.** What is an electromagnetic wave?

**Q1.22.** What are macroscopic quantities?

# PROBLEMS

**P1.1.** How many electrons are needed to obtain one coulomb (1 C) of negative charge? Compare this number with the number of people on earth (about $5 \cdot 10^9$).

**P1.2.** Calculate approximately the gravitational force between two glasses of water a distance $d = 1$ m apart, containing 2 dl (0.2 liter) of water each.

**P1.3.** Estimate the amount of equal negative electric charge (in coulombs) in the two glasses of water in problem P1.2 that would cancel the gravitational force.

**P1.4.** Two small equally charged bodies of masses $m = 1$ g are placed one above the other at a distance $d = 10$ cm. How much negative charge would the bodies need to have so that the electric force on the upper body is equal to the gravitational force on it (i.e., so the upper body levitates)? Do you think this charge can be realized?

**P1.5.** Calculate the electric field strength necessary to make a droplet of water of radius $a = 10\,\mu$m, with an excess charge of 1000 electrons, levitate in the gravitational field of the earth.

**P1.6.** How large does the electric field intensity need to be in order to levitate a body 1 kg in mass and charged with $-10^{-8}$ C? Is the answer of practical value, and why?

**P1.7.** A drop of oil, $r = 2.25\,\mu$m in radius, is negatively charged and is floating above a very large, also negatively charged body. The electric field intensity of the large body happens to be $E = 7.83 \cdot 10^4$ V/m at the point where the oil drop is situated. The density of oil is $\rho_m = 0.851$ g/cm³. (1) What is the charge of the drop equal to? (2) How large is this charge compared to the charge of an electron? *Note: the values given in this problem can realistically be achieved in the lab. Millikan used such an experiment at the beginning of the 20th century to show that charge is quantized.*

**P1.8.** Find the force between the two parallel wire segments in Fig. P1.8 if they are 1 mm long and 10 cm apart, and if they are parts of current loops that carry 1 A of current each. The constant $k_m$ is equal to $10^{-7}$ in SI units (N/A²).
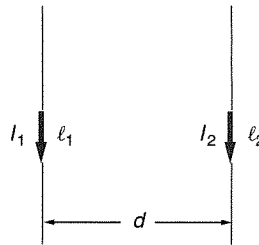


Figure P1.8 Two parallel wire segments

**P1.9.** A small body charged with $Q = -10^{-10}$ C finds itself in a uniform electric and magnetic field as shown in Fig. P1.9. The electric field vector and the magnetic flux density vector are **E** and **B**, respectively, everywhere around the body. If the magnitude of the electric field is $E = 100$ N/C, and the magnetic flux density magnitude is $B = 10^{-4}$ N $\cdot$ s/C $\cdot$ m, find the force on the body if it is moving with a velocity **v** as shown in the figure, where $v = 10$ m/s (the speed of a slow car on a mountain road). How fast would the body need to move to maintain its direction of motion?
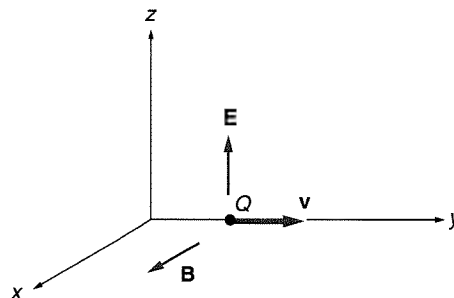


Figure P1.9 Point charge in an electric and magnetic field

**P1.10.** Volta used a chemical reaction to make the first battery that could produce continuous electric current. Use the library, or any other means, to find out if electric current can be used to make chemical reactions possible. Write one page on the history and implications of these processes.

# 2

# Circuit Theory and Electromagnetics

## 2.1    Introduction

One of the most important tools of electrical engineers is circuit theory. Circuits have charges and currents, which we know produce electric and magnetic fields. Thus circuits are actually electromagnetic systems and strictly speaking, they should be analyzed starting from the general electromagnetic-field equations, i.e., Maxwell's equations. We start, however, from the two Kirchhoff's laws instead, well aware that circuit theory can be used to accurately predict circuit behavior.

Circuit theory is an approximate theory that can be obtained from Maxwell's equations with a set of approximations. We will return to this point throughout this book. In this chapter we review some simple circuit examples and look at where these approximations are made, arriving at two important conclusions. First, circuit theory is an approximation (but fortunately a very good and useful one in most applications). Second, the limitations of circuit theory can be understood only if we understand electromagnetic-field theory. In the next section we consider the effects of some simple electromagnetic properties of electric circuits, which will help you understand these conclusions. You can perform the examples shown here in the lab, using just a function generator and oscilloscope.
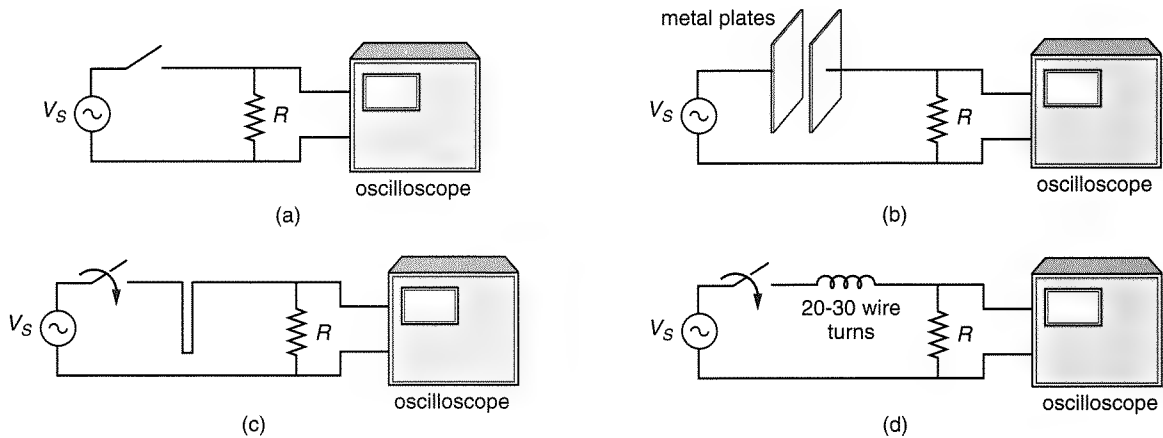
Figure 2.1 (a) A resistor connected to a function generator. The voltage across the resistor is observed on an oscilloscope. (b) The switch in (a) is replaced by two metal plates. (c) An interconnecting conductor is shaped to form a short-circuited two-wire line. (d) The interconnecting conductor is wound around a pen.

## 2.2    Circuit Elements as Electromagnetic Structures

Let us first consider three basic circuit elements: a switch, a conducting wire, and a resistor (Fig. 2.1). We shall find out later how Ohm's law is derived from Maxwell's equations, but right now let us start from what you learned in circuits: the voltage across the resistor is $v_R(t) = Ri(t)$. (Actually, this relation should be considered as the definition of an *ideal* resistor.)

We connect the resistor to a function generator and look at the voltage across it on an oscilloscope, Fig. 2.1a, when the switch is open. The classical expectation is that the voltage is zero. However, note that the switch consists of two contacts that are separated by an insulator (e.g., air) when the switch is open. These two contacts, therefore, form a capacitor. Only if we can neglect the capacitance of this capacitor, i.e., if we can consider it to be zero, is the voltage across the resistor also zero.

To understand this, imagine changing the *shape* and *size* of the switch. For example, let us replace the actual switch by two parallel rectangular metal plates, say 10 by 10 cm. Let the plates be separated by 1 cm when the switch is open, as in Fig. 2.1b, and pressed tightly together when the switch is closed. The resistance of the large plates is certainly less than that of the switch contacts, and close to $0\,\Omega$, so we should see no change in $v_R(t)$ on the oscilloscope screen. However, in the open position this new switch may influence the current in the circuit considerably because it has a sizable capacitance. Indeed, if we bring the two plates closer together, we will notice on the oscilloscope that this open switch influences the voltage between the resistor terminals more than when the plates are farther apart.

This capacitance is present in any switch, but if—as mentioned—the capacitance is small enough, the switch will *behave* as if the capacitance is zero. However, in strict electromagnetic theory even an element as simple as a switch does not exist. Only if we can neglect its capacitance does a switch behave according to the defi-

nition in circuit theory, i.e., that it is either an open switch or a short circuit. To analyze the open switch more accurately, we must consider its capacitance and use electromagnetic-field theory. Recall that the reactance of a capacitor is inversely proportional to the product of its capacitance and frequency. So we may infer that, at extremely high frequencies, it may not be easy to make a switch that, if open, acts indeed as an open circuit.

Let us now concentrate on the influence of the size and shape of a conducting wire. In circuit theory, the wire form and size are assumed to have no effect on circuit behavior and they are assumed to be short-circuit interconnections. We now analyze this assumption in more detail.

Assume that the switch is closed. According to circuit theory, we may vary the length and shape of the wire connecting the resistor to the generator as much as we wish without changing either the voltage across the resistor or current in it. What will happen, however, if we substantially extend one of the conductors and bend it as in Fig. 2.1c, so that we get two relatively long parallel close wires? Experiments show that this changes the voltage across the resistor to a large degree. How can we explain this?

The bent conductor represents a section of short-circuited transmission line. If the frequency is low enough, this is just a wire loop having a certain inductance. This inductance is connected in series with the resistor and changes the current, and hence the voltage across the resistor.

So the circuit-theory assumption that the interconnecting conductors have no effect on the circuit behavior is only an approximation. Electromagnetic-field theory tells us that in the case of varying currents, the same circuit will behave differently if we twist it, extending, shortening, or deforming the interconnecting conductors and generally changing the circuit's shape. This is indeed a strange conclusion if one adheres to circuit-theory explanations, but it is true. At high frequencies, even as low as about 10 MHz, and for circuit dimensions exceeding about 10 cm, circuit theory frequently cannot predict circuit properties with sufficient accuracy, but electromagnetic theory can.

As a more specific example, consider the circuit in Fig. 2.2. It consists of one resistor and one capacitor of very small dimensions (known as "chip" or "surface mount" resistors and capacitors). If we compute the input impedance of the circuit as a function of frequency, we get the solid line in Fig. 2.3. Experimentally obtained results, indicated by the square symbols, are quite different, however. Above a certain
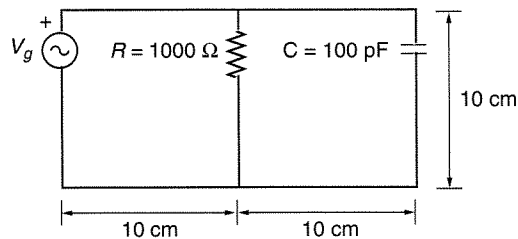


Figure 2.2 A simple circuit with a small resistor and a small capacitor
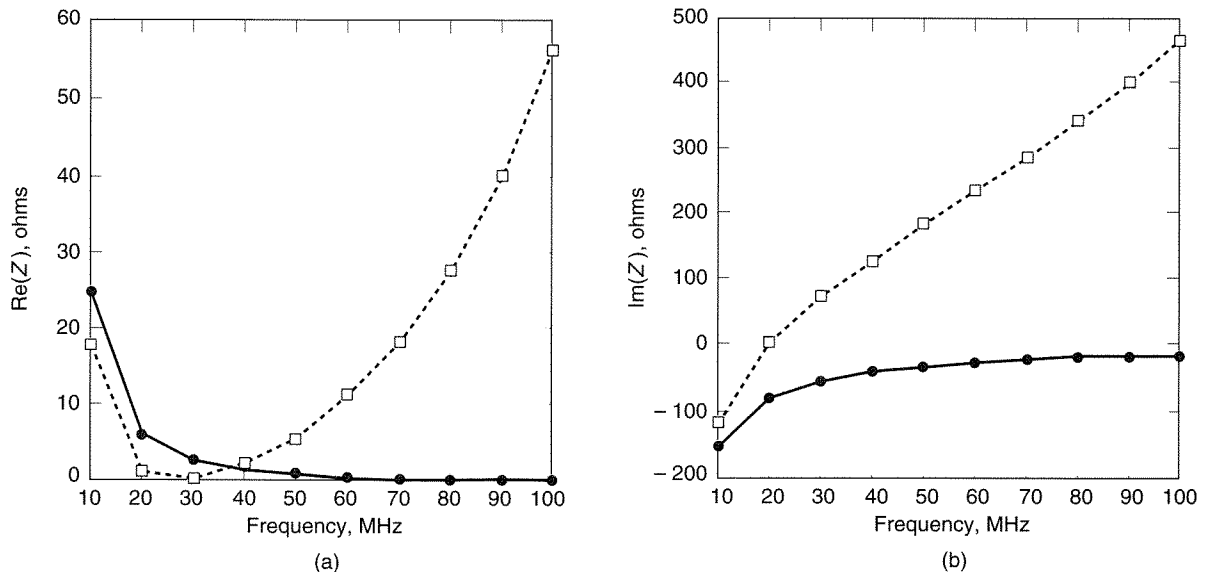
**Figure 2.3** Real (a) and imaginary (b) parts of the input impedance of the circuit shown in Fig. 2.2 versus frequency, obtained by circuit theory (solid line with dots), by electromagnetic analysis of the circuit (dashed line), and by experiment (small squares).

frequency, they differ greatly from those predicted by circuit theory. We now know why: in addition to the circuit elements themselves (the resistor and the capacitor), the shape of interconnecting wires in Fig. 2.2 also influences the circuit behavior. This simple circuit can also be analyzed using electromagnetic theory and computer programs that take the shape of the interconnecting conductors into account. The result for the circuit impedance using such a program is shown in Fig. 2.3 in dashed line. You can observe excellent agreement between measurement and theory at frequencies considerably above those where circuit theory loses accuracy. For the moment, you may trust (or not trust) the dashed line results.

Another effect observed in the circuit from Fig. 2.2 is associated with the assumption that the chip (surface mount) components are very small, or *lumped* (which is always assumed in circuit theory). The chip capacitor and resistor in Fig. 2.2 will in reality not have the exact impedance values given in their specification sheets. It turns out that most chip capacitors and resistors have an associated series lead inductance of about 1 nH. That means that above a certain frequency, the chip capacitor will start behaving like an inductor. It is left as an exercise for the reader to calculate this resonant frequency for 1, 10, and 100-pF chip capacitors.

As the next example, let us again close the switch in Fig. 2.1a. Then we take the wire connecting the resistor to the generator and wind it tightly around a pen 20 to 30 times, as shown in Fig. 2.1d. The result is a more conventional inductor than the bent conductor of Fig. 2.1c. Again the shape of the wire has a huge effect on the voltage across the resistor. It has an even larger effect if an iron rod is used instead of the pen. We can find the value of inductance only by using electromagnetic theory, or by measurements. What we will learn in this book is how to take into account the actual
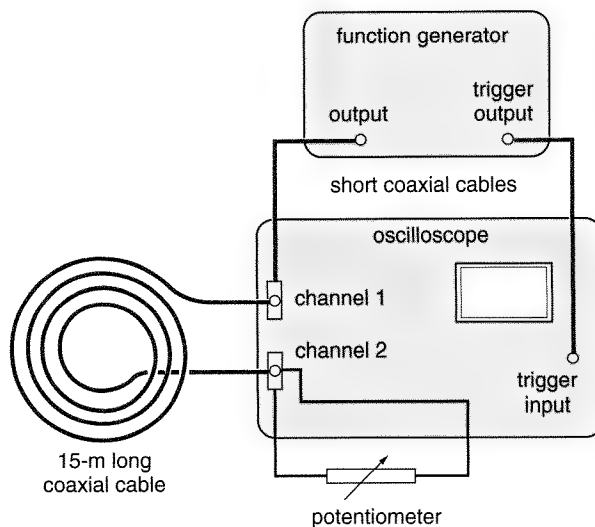
**Figure 2.4** Observing electromagnetic effects in a coaxial cable

shape of conductive and nonconductive bodies, and the properties of the materials they are made of, in order to predict the behavior not only of simple circuits but also of different devices used in electrical engineering.

The last example is related to electromagnetic waves and transmission-line theory. Figure 2.4 shows a 15-m coaxial cable connected at one end to a function generator. At the other end it is connected to channel 2 of a two-channel oscilloscope and in parallel with a potentiometer (variable resistor). Channel 1 of the oscilloscope monitors the output of the function generator (which is the input to the coaxial cable). If you used only basic circuit theory, you would expect to see the same voltage for all values of the resistor at the end of the cable, and the voltage should be the same as that coming out of the signal generator. However, due to electromagnetic wave effects, the waveforms at the two channels (the voltages at the beginning and the end of the coaxial cable) can be very different. For example, a 1-V pulse from the signal generator could result in a negative, zero, or greater-than-1-V pulse at channel 2 of the oscilloscope. To explain this result we need so-called transmission-line theory, which turns out to be a special case of electromagnetic wave theory. We shall consider transmission lines in Chapter 18.

*Questions and problems*    Q2.1 to Q2.7, P2.1 to P2.4

## 2.3     Oscillations in Circuits from the Electromagnetic Point of View

Let us review a more complex (but still very simple) circuit shown in Fig. 2.5, which is a combination of the previous cases. This is a series resonant circuit. For a voltage
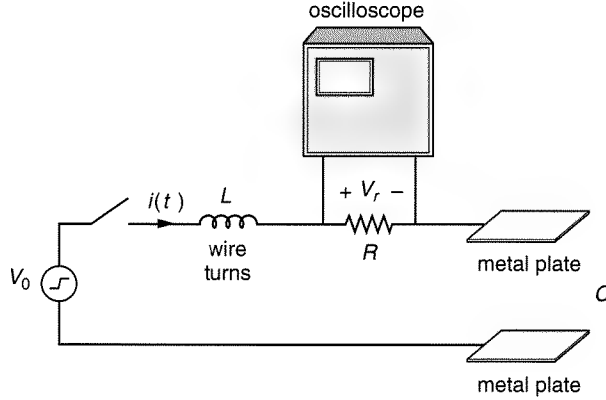
Figure 2.5 A possible physical realization of a series resonant circuit

step $V_0$ turned on at $t = 0$, Kirchhoff's voltage law gives

$$Ri + L\frac{di}{dt} + \frac{1}{C}\int_0^t i\,dt + V_0 = 0. \tag{2.1}$$

By differentiating with respect to $t$ and rearranging the terms, we get

$$\frac{d^2i}{dt^2} + \frac{R}{L}\frac{di}{dt} + \frac{i}{LC} = 0, \tag{2.2}$$

which is a second-order ordinary differential equation with exponential solutions of the form

$$i(t) = A_1 \exp(s_1 t) + A_2 \exp(s_2 t). \tag{2.3}$$

$A_1$ and $A_2$ are constants determined from the initial conditions, and $s_1$ and $s_2$ are complex roots of the characteristic equation of (2.2), given by

$$s_{1,2} = -\frac{R}{2L} \pm \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC}}. \tag{2.4}$$

Let $\omega_0 = 1/\sqrt{LC}$. For $\omega_0^2 > (R/2L)^2$, $s_{1,2} = -\alpha \pm j\omega$ are complex numbers with real and imaginary parts, as seen in Eq. (2.4), and the solution for the current is

$$i(t) = B_1 e^{-\alpha t}\cos\omega t + B_2 e^{-\alpha t}\sin\omega t. \tag{2.5}$$

The constants $B_{1,2}$ are given by the initial conditions. At $t = 0$ there is no current through the inductor before the voltage is turned on at the input, $i(0) = 0$ and $B_1 = 0$, so $i(t) = B_2 e^{-\alpha t}\sin\omega t$, and $B_2$ can be found from knowing what $di(0)/dt$ is. Since $i(0) = 0$ immediately after the switch is closed, there is no voltage drop across the resistor and the initial voltage $V_0$ on the capacitor shows up across the inductor,

$L(di/dt)_{t=0} = V_0$. The final expression for the voltage across the resistor after the switch is closed is

$$v_R(t) = Ri(t) = R\frac{V_0}{L}e^{-\alpha t}\sin\omega t. \tag{2.6}$$

This last expression shows that the voltage is a sinusoid with an exponential amplitude decay. This is called a *damped oscillation*. In electromagnetic terms, the energy in an undamped case is stored in the inductor for one half of the cycle, and in the capacitor in the other half. In a damped case, some of the electromagnetic energy goes into heat in the resistive parts of the circuit.

This effect can also be explained in a similar way by circuit theory. What *cannot* be answered by circuit theory, however, are the following questions:

- Does the resonant frequency depend on the circuit shape and size?
- Does the damping depend on the circuit shape and size?

We already know the answer to the first question: the resonant frequency *does* depend (at least to some extent) on the shape of the circuit. The reasoning is exactly as in the previous examples.

The second question itself seems a bit strange: how can we have more damping than that resulting from losses in the resistor? We mentioned in the first chapter that Maxwell predicted the existence of electromagnetic waves. We will learn that theoretically these waves are produced by *all* systems with time-varying currents, and that the efficiency in producing these waves depends on the system size and shape. We will also learn that an electromagnetic wave is, in fact, an energy package. Thus, "radiation" of electromagnetic waves actually implies leakage, or loss, of energy from the system producing them. Therefore resonant circuits *do* have damping that depends on their size and shape. Fortunately, in most applications this effect is negligible, but it always exists. It can be predicted only by electromagnetic-field theory—circuit theory is unable to do that. We will learn how large a circuit must be to radiate substantially.

*Questions and problems*    Q2.8 and Q2.9, P2.5 to P2.8

## 2.4    Chapter Summary

1. Circuit theory is not exact; it is an approximation of electromagnetic-field theory.
2. To understand the limitations of circuit theory, we have to begin from electromagnetic-field theory.
3. To determine theoretically the capacitance of a capacitor or the inductance of a coil, it is necessary to use electromagnetic-field theory. The calculation of the resistance of a resistor also requires some knowledge of electromagnetic-field theory.

4. Along transmission lines, such as two-wire or coaxial lines, exist specific electromagnetic waves with specific effects that cannot be explained in terms of circuit-theory concepts.

5. Resonance effects and damping in circuits depend on the circuit shape and size, a strange phenomenon from the circuit-theory viewpoint.

# QUESTIONS

**Q2.1.** Why does every switch have capacitance?

**Q2.2.** Try to imagine a "perfect" but real switch (a switch with the smallest possible capacitance). How would you design a good switch? What would be its likely limitations?

**Q2.3.** Why does it become progressively more difficult to have an "ideal" switch as frequency increases?

**Q2.4.** Why is the circuit-theory assumption that interconnecting conductors (wires) have no effect on the circuit behavior incorrect?

**Q2.5.** Imagine a resistor connected to a car battery by wires of fixed length. Does the shape of the wires influence the current in the resistor? Explain.

**Q2.6.** Answer question Q2.5 if the source is a (1) 60-Hz and (2) 1-GHz generator.

**Q2.7.** Give at least two reasons for the failure of circuit theory when analyzing the simple circuit in Fig. 2.2.

**Q2.8.** Explain why the resonant frequency of a circuit is at least to some extent dependent on the circuit shape and size.

**Q2.9.** Why does the damping in resonant circuits depend at least to some extent on the circuits' shape and size? Can circuit theory explain this?

# PROBLEMS

**P2.1.** The capacitance of a switch ranges from a fraction of a picofarad to a few picofarads. Assume that a generator of variable angular frequency $\omega$ is connected to a resistor of resistance of $1\,M\Omega$, but that the switch is open. Assuming a switch capacitance of $1\,pF$, at what frequency is the open switch reactance equal to the resistor resistance?

**P2.2.** A surface-mount capacitor has a 1-nH parasitic series lead inductance. Calculate and plot the frequency at which such a capacitor starts looking like an inductor, as a function of the capacitance value.

**P2.3.** A surface-mount resistor of resistance $R = 100\,\Omega$ has a 1-nH series lead inductance. Plot the real and imaginary part of the impedance as a function of frequency. In which frequency range can this chip be used as a resistor?

**P2.4.** The windings of a coil have a parasitic capacitance of 0.1 pF, which can be viewed as an equivalent ~~series~~ capacitance. Plot the reactance of such a $1\,\mu H$ coil as a function of frequency.    *Parallel*

**P2.5.** A capacitor of capacitance $C$ receives a charge $Q$. It is then connected to an uncharged capacitor of the same capacitance $C$ by means of conductors with practically no resistance. Find the energy contained in the capacitor before connecting it to the other

capacitor, and the energy contained in the two capacitors. [The energy of a capacitor is given by $W_e = Q^2/(2C)$]. Can you explain the results using circuit-theory arguments? Can you explain the results at all?

**P2.6.** The inductance of a thin circular loop of radius $R$, made of wire of radius $a$, where $R \gg a$, is given by the approximate formula

$$L_0 \simeq \mu_0 R \left( \ln \frac{8R}{a} - 2 \right) \text{(henries)},$$

where $\mu_0 = 4\pi \, 10^{-7}$ H/m, and $R$ and $a$ are in meters. A capacitor of capacitance $C = 100$ pF and a coil of inductance $L = 100$ nH are connected in series by wires of radius $a$ and the shape of a circular loop of radius $R$ ($R \gg a$). Find: (1) the radius of the loop that results in $L_0 = L$ if $a = 0.1$ mm; (2) the radius of the wire that results in $L_0 = L$ if $R = 2.5$ cm; and (3) the resonant frequency of the circuit versus the loop radius, $R$, if $a = 0.1$ mm.

**P2.7.** The capacitance between the terminals of a resistor is $C = 0.5$ pF, and its resistance is $R = 10^6 \, \Omega$. Plot the real and imaginary part of the impedance of this dominantly resistive element versus frequency from 0 MHz to 10 MHz.

**P2.8.** A coil is made in the form of $N = 10$ tightly packed turns of wire. Predict qualitatively the high-frequency behavior of the coil. Explain your reasoning.

# 3

# Coulomb's Law in Vector Form and Electric Field Strength

## 3.1 Introduction

We have seen that sources of an electrostatic field are stationary and time-constant electric charges. This is the simplest form of the general electromagnetic field: since there are no moving or time-varying charges, there is no magnetic field. Although the electrostatic field is only a special case of the electromagnetic field, it occurs frequently. It is essentially the field that drives the electric current through wires and resistors in electric circuits; the field driving tiny signals in our nerves and brain; the field in depletion layers of transistors in computer chips; or the field that ionizes air (makes it conducting) just before a lightning bolt.

In this introductory course, the electrostatic field is of specific importance. The simplicity of the physical concepts in the electrostatic field allows us to develop mathematical models in a straightforward way. Later on, in more complex fields, we will be able to solve difficult practical problems using these concepts and tools as they are or with minor modifications.

28

## 3.2    Coulomb's Law in Vector Form

We have seen that Coulomb's law is an experimentally established law which describes the force between two charged bodies that are small compared to the distance between them. Such charged bodies are referred to as point charges. Coulomb's law in Eq. (1.1) is an algebraic expression that needs an additional explanation in words. The force directed along the line joining the two bodies is either repulsive or attractive. It is repulsive if the two charges are of the same kind, or sign, and attractive if they are of different kind.

Using vector notation, it is not difficult to write Coulomb's law in a form that does not need such additional explanations. Let $\mathbf{r}_{12}$ be the vector directed from charge $Q_1$ to charge $Q_2$ (Fig. 3.1), and $\mathbf{u}_{r12} = \mathbf{r}_{12}/|\mathbf{r}_{12}|$ be the unit vector of $\mathbf{r}_{12}$. (A number of notations have been used for unit vectors, e.g., $\mathbf{a}_r$, $\mathbf{u}_r$, or $\hat{\mathbf{r}}$ for the unit vector of vector $\mathbf{r}$. We will adopt $\mathbf{u}_r$ to remind us that it is a *unit* vector.) We can then express Coulomb's law in Eq. (1.1) in the following form:

$$\mathbf{F}_{e12} = \frac{1}{4\pi\epsilon_0} \frac{Q_1 Q_2}{r^2} \mathbf{u}_{r12} \qquad \text{newtons (N).} \qquad (3.1)$$

*(Coulomb's law in vector form)*

The unit vector $\mathbf{u}_{r12}$, and thus also the force $\mathbf{F}_{e12}$, is directed from $Q_1$ toward $Q_2$, so that the additional explanation in words is not necessary anymore. Moreover, we have adopted the convention that a positive charge implies a *positive sign*, and a negative charge a *negative sign*. Complete information about the direction of the force is therefore also contained in this vector expression (recall that $-\mathbf{r}$ means the same vector, but in the opposite direction). If the two charges are of the same sign, the vector $\mathbf{u}_{r12}$ is as in Fig. 3.1, and if they are of different signs, we have instead $-\mathbf{u}_{r12}$, which means that the force is attractive. (If necessary, before proceeding further read Sections A1.1–A1.3 of Appendix 1, "Brief survey of vectors and vector calculus.")

The constant $\epsilon_0$ is known as the *permittivity of free space* or *of a vacuum*. According to Eq. (3.1) its unit is $C^2/(m^2\,N)$. Usually a simpler unit, F/m (farads per meter), is used, as will be explained in Chapter 8. The value of $\epsilon_0$ is
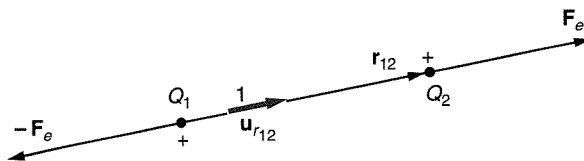
Figure 3.1 Notation in the vector form of Coulomb's law
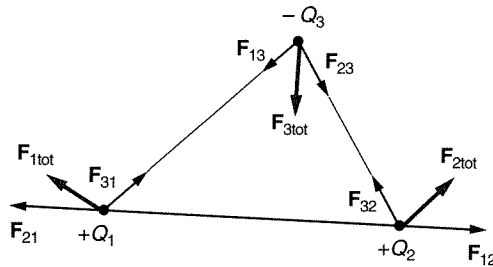
Figure 3.2 Example of vector addition of
Coulomb forces

$$\epsilon_0 = 8.854 \cdot 10^{-12} \text{ farads per meter (F/m)} \approx \frac{1}{36\pi \cdot 10^9} \text{ F/m.} \tag{3.2}$$

$$(Permittivity\ of\ a\ vacuum)$$

The reason for writing $1/(4\pi\epsilon_0)$ in Coulomb's law instead of, say, simply $k$, is purely practical, removing the factor $4\pi$ from many other commonly used equations. Also, in many equations the permittivity of free space then appears as it is, $\epsilon_0$, and not as its reciprocal.

Coulomb measured the electric force in air. We will see later that the electrical properties of air are very nearly the same as those of a vacuum, i.e., of space with no elementary particles of matter. Coulomb's law in Eq. (3.1) is therefore valid for charges that are strictly in a vacuum, but the presence of air does not change the result substantially. Therefore, the term "free space" usually implies vacuum or air.

Coulomb also measured the force on one point charge (e.g., $Q$) due to several point charges (e.g., $Q_1, Q_2, \ldots$). He concluded that the total force is obtained by *vector addition* of Coulomb's forces acting on $Q$ by charges $Q_1, Q_2, \ldots$ individually. We know that mechanical forces on a body are summed in the same way, which is known as the *principle of superposition for forces*. An example of vector addition of Coulomb forces is sketched in Fig. 3.2.

How large are charges and electric forces we encounter around us? The charges rarely exceed a few nanocoulombs ($1\,\text{nC} = 10^{-9}\,\text{C}$). The largest electric forces around us do not exceed about $1\,\text{N}$, which is the weight of a small glass of water. Thus, measured by our standards, electric forces are very small.

*Questions and problems:*   Q3.1 to Q3.10, P3.1 to P3.11

## 3.3    Electric Field Strength of Known Distribution of Point Charges

Let us repeat the definition of the electric field strength vector given in Eq. (1.5) in somewhat different notation. We first define the *test charge*, $\Delta Q$, to be a very small

body with negligibly small charge. (Such a test charge placed in an electric field will not affect the field, so that we can measure the electric field strength at a particular point of the field as it is in the absence of the test charge.) Then at any point in the electric field, the electric field strength vector is given by

$$E = \frac{F_{on\ \Delta Q}}{\Delta Q} \qquad \text{newtons per coulomb (N/C)} = \text{volts per meter (V/m)}. \quad (3.3)$$

*(Definition of the electric field strength vector)*

The unit of the electric field strength is newtons per coulomb (N/C). For reasons to become clear in the next chapter, the equivalent unit, volts per meter (V/m), is used instead.

Combining this definition with the expression for the force exerted by one point charge on another point charge, Eq. (3.1), we see that the electric field vector due to a point charge $Q$ is given by

$$E = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} u_r \qquad (V/m), \qquad\qquad (3.4)$$

*(Electric field strength of a point charge)*

where $u_r$ is the unit vector directed *away* from charge $Q$. This is the electric field strength of a single point charge. It is a vector function of the distance from the point charge producing it and is directed away from a positive charge or toward a negative charge.

What if we have more than one charge producing the field? How is the electric field strength then obtained? The answer is fairly obvious: since the principle of superposition is valid for electric forces just as for mechanical forces, the equation follows directly from Eqs. (3.3) and (3.4). Assume that we have $n$ point charges, $Q_1, Q_2, \ldots, Q_n$. The electric field strength at a point that is at distances $r_1, r_2, \ldots, r_n$ from the charges is simply

$$E = \sum_{i=1}^{n} \frac{1}{4\pi\epsilon_0} \frac{Q_i}{r_i^2} u_{ri} \qquad (V/m). \qquad\qquad (3.5)$$

**Example 3.1—Superposition applied to the electric field strength.** As an example, Fig. 3.3 shows how we obtain the electric field strength resulting from three point charges, $Q, 2Q$, and $3Q$. Assume that the three charges are in air, in the plane of the drawing, and let us determine the total electric field at the point $P$ which is at the same distance from the three charges. To obtain the total field, we first add up the field of the charges $Q$ and $2Q$, and then add to this sum the field of the charge $3Q$, as indicated in the figure.

It is important to note that Eq. (3.5) can be used in this case because both vacuum and air, in which the charges are placed, are *linear media* (i.e., electrical properties of the medium, in this case of $\epsilon_0$, do not depend on the electric field strength in the medium). This is not always the
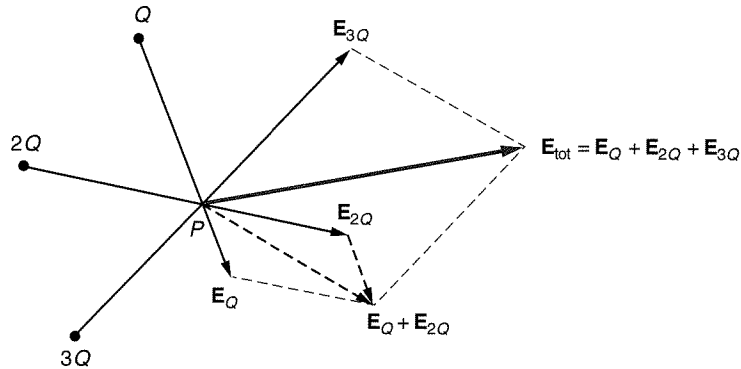
Figure 3.3 The electric field strength resulting from the three point charges $Q$, $2Q$, and $3Q$, situated in the plane of the drawing, at a point that is at the same distance from each of them

case—many important and practical media are nonlinear. For nonlinear media, superposition cannot be applied. Superposition allows us to break up a complicated problem into several easier ones and then add up their solutions to get the solution to the complicated problem. We will use it often.

*Questions and problems:*   Q3.11 to Q3.14, P3.12 to P3.15

## 3.4    Electric Field Strength of Volume, Surface, and Line Charge Distributions

Elemental charges (electrons and protons) that create a field are always so small that they can be considered as point charges. Therefore Eq. (3.5) can be used, in principle, to calculate the electric field of any charge distribution. However, even for a tiny amount of charge, the number of elemental charges is very large. For example, $-1\,\mathrm{pC}$ ($-10^{-12}\,\mathrm{C}$) contains about $10^7$ electrons. Therefore, in macroscopic electromagnetism, it is convenient to introduce the concept of *charge density*, and then to use integral calculus to evaluate the field of a charge distribution.

### 3.4.1    VOLUME CHARGE DENSITY

Consider first a cloud of static charges. (Of course, it must be kept in place by some means; otherwise it would move as a result of the electric forces the charges exert on each other.) Let them be packed so densely that even inside a small volume $dv$ there are many charges, amounting to a total charge $dQ_{\text{in }dv}$. We then define the *volume charge density*, $\rho$, at the point enclosed by that small volume:

$$\rho = \frac{dQ_{\text{in }dv}}{dv} \qquad \text{coulombs per cubic meter (C/m}^3\text{).} \tag{3.6}$$

*(Definition of volume charge density)*

Note that the unit of volume charge density is $\mathrm{C/m}^3$.
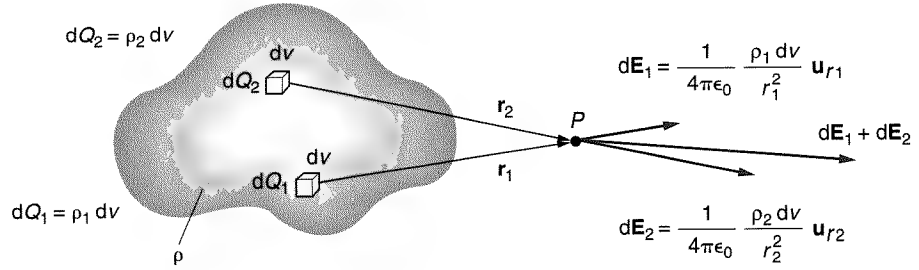
Figure 3.4 Calculating the electric field strength of a charged cloud of known charge density

According to this definition, inside a small volume $dv$ where the charge density is $\rho$, there is a small charge

$$dQ = \rho \, dv. \tag{3.7}$$

This charge can be considered a point charge. If we have a charged cloud with known charge density $\rho$ at all points, we can obtain the electric field strength at any point using, essentially, Eq. (3.5) but with a very large (theoretically infinitely large) number $n$ of point charges. Such a sum is an integral (see Fig. 3.4):

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \int_v \frac{\rho \, dv}{r^2} \mathbf{u}_r \quad \text{(V/m)}. \tag{3.8}$$

*(Electric field strength of volume distribution of charges)*

Note that the charge density, $\rho$, and the ~~position vector, r,~~ vary from one elemental volume $dv$ to another.    DISTANCE $r$ FROM THE FIELD POINT & $a_r$

Suppose we know the shape of the cloud and volume charge density at all points of the cloud. It is possible only rarely to evaluate the integral in Eq. (3.8) analytically. However, we can approximately calculate the electric field strength at any point in space by dividing the volume charge into a finite number of very small volumes $\Delta v$, taking the value of $\rho$ at the center of this small volume, and then summing all the vector electric field strengths resulting from these point charges. In this case, Eq. (3.8) becomes a sum over all the little volumes. This sum is not hard to evaluate on a computer, and the result will be more accurate with a greater number of small volumes.

## 3.4.2 SURFACE CHARGE DENSITY

Strictly speaking, volume charge is the only type of charge that appears in nature. However, in some cases this charge is spread in an extremely thin layer (of thickness on the order of a few atomic radii) and can be regarded as a *surface charge*. Such is, for example, the excess charge on a conducting body. To describe the surface charge distribution, we introduce the concept of *surface charge density*, $\sigma$. Figure 3.5 shows a body with surface charge. Consider a small area $dS$ of the body surface. Let the charge on that small surface patch be $dQ_{\text{on } dS}$. The surface charge density at a point

$$dQ = -\sigma\, dS$$

$$dE = \frac{1}{4\pi\epsilon_0}\frac{-\sigma\, dS}{r^2}\mathbf{u}_r$$
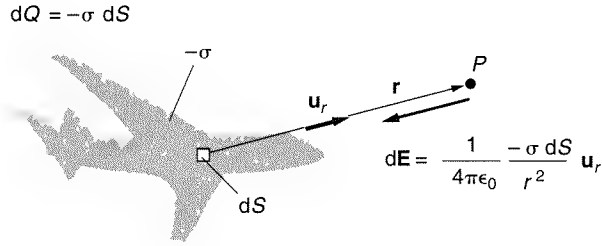
Figure 3.5 A body charged over its surface

of $dS$ is then defined as

$$\sigma = \frac{dQ_{on\ dS}}{dS} \qquad \text{coulombs per square meter (C/m}^2). \tag{3.9}$$

*(Definition of surface charge density)*

(The symbol $\rho_s$ is sometimes used instead of $\sigma$.) From this definition, it follows that if we know the surface charge density at a point on the body surface, the charge on a small patch of area $dS$ enclosing this point is obtained as

$$dQ_{on\ dS} = \sigma\, dS. \tag{3.10}$$

The unit of surface charge density is $C/m^2$. Note that in general, the surface charge differs from one point of a surface to another.

The field of a given distribution of surface charge is obtained if in Eq. (3.8) we substitute the elemental charge, $\rho\, dv$, by $\sigma\, dS$:

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\int_S \frac{\sigma\, dS}{r^2}\mathbf{u}_r \qquad \text{(V/m)}. \tag{3.11}$$

*(Electric field strength of a surface distribution of charges)*

### 3.4.3   LINE CHARGE DENSITY

Finally, we frequently encounter thin charged wires. Wires are usually conductors and the charge is distributed in a very thin layer on the wire surface. If the wire is thin compared to the distance of the observation point, we can consider the charge to be distributed approximately along a geometric line, for example along the wire axis. This type of charge is known as *line charge*. Its distribution along the line (i.e., along the wire the line approximates) is described by the *line charge density*, $Q'$ (Fig. 3.6).

Let the charge on a very short segment $dl$ of the line be $dQ_{on\ dl}$. The line charge density is defined as

$$Q' = \frac{dQ_{on\ dl}}{dl} \qquad \text{coulombs per meter (C/m)}. \tag{3.12}$$

*(Definition of line charge density)*

(The symbol $\rho_\ell$ is sometimes used instead of $Q'$.) Thus if we know the line charge density at a point along a wire, the charge on a short segment $dl$ of the wire containing
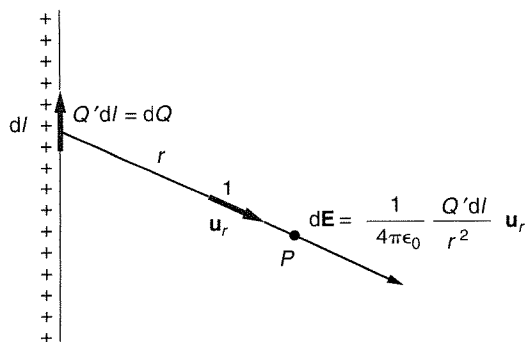
**Figure 3.6** Electric field strength due to a thin charged wire

that point is simply

$$dQ_{\text{on } dl} = Q' \, dl. \tag{3.13}$$

The unit of line charge density is C/m. Note that $Q'$ may differ from one point of the line to another.

The field of a given distribution of line charge along a line $L$ is obtained as

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \int_L \frac{Q' \, dl}{r^2} \mathbf{u}_r \qquad \text{(V/m)}. \tag{3.14}$$

*(Electric field strength of line distribution of charges)*

The integrals in Eqs. (3.8), (3.11), and (3.14) usually cannot be evaluated analytically, but they can always be evaluated numerically, as described in connection with the volume charge distribution. We do not give any examples of analytical evaluation of these integrals because for those that can be evaluated, it is usually possible to obtain the result in a much simpler way, described in the next two chapters.

If we know the distribution of volume, surface, and line charges, i.e., if we know the volume, surface, and line charge density at all points, it is a simple matter to evaluate the electric field strength of these distributions at any point. As we shall see, however, the charge distribution is rarely known in advance. Instead, in practical problems we need to *determine* the charge distribution in order to calculate the electric field around it. Therefore, the formulas to determine the electric field strength from a known distribution of volume, surface, or line charges are mainly of academic interest. The concepts of volume, surface, and line charge are very useful, however, as we shall see later. Equations in the form of Eqs. (3.11) and (3.14) are also indispensable in determining unknown charge distributions numerically.

*Questions and problems:* Q3.15 to Q3.18, P3.16 to P3.27

## 3.5 Lines of the Electric Field Strength Vector

The electrostatic field is a vector field. A useful concept for visualizing vector fields is *field lines*. The lines of vector **E** are defined as imaginary, generally curved lines,
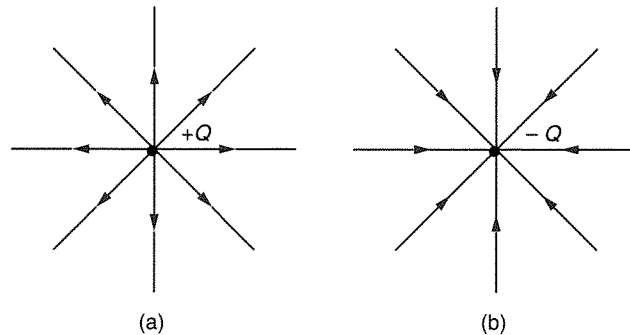
Figure 3.7 Electric field lines of (a) a positive and (b) a negative point charge

having the property that **E** is tangential to these lines at all points. For example, lines of vector **E** of the field of a point charge are straight lines emanating from the charge (Fig. 3.7). It is usual to add an arrow to the lines of vector **E** indicating the direction of **E** along the lines.

**Example 3.2—Electric field lines of a very large, uniformly charged plate.** As a further example of electric field lines, consider a very large, uniformly charged flat plate. Let the charge on the plate be positive. Since the plate is very large, if we consider the field close to the plate, the electric field strength vector must be normal to the plate and pointing away from it (why?). The electric field lines are as sketched in Fig. 3.8. This kind of electric field, which has in a region of space the same direction and magnitude, is called a *uniform electric field.*

For a negative plate, the lines are of the same form, only the arrowhead (indicating the direction of vector **E**) points toward the plate, i.e., the field is also uniform.
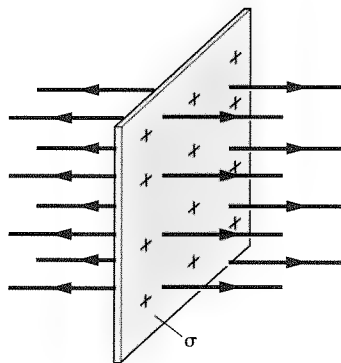


Figure 3.8 Electric field lines for a flat, large plate with uniform positive surface charge distribution

## 3.6    Chapter Summary

1. If written in vector form, Coulomb's law does not require additional explanations in words—all information is then contained in the formula. For formulating Coulomb's law in this manner, the positive and negative sign convention for charges is essential.

2. The electric field strength vector **E** of a point charge is defined from the vector form of Coulomb's law.

3. The expression for the electric field strength of a point charge can be used for obtaining vector **E** resulting from any distribution of point charges, or from volume, surface, and line distributions of charges. To do this, it is necessary to define volume, surface, and line charge distributions.

4. The charge distribution is usually not known in advance. Therefore, the formulas for the electric field strengths of known distributions of charges are of limited practical usefulness. However, they may be used for numerical evaluation of the vector **E** by means of integral equations.

5. We say that in a region of space the electric field is uniform if the electric field strength at all points of the region has the same direction and magnitude.

## QUESTIONS

**Q3.1.** Discuss the statement that Eq. (3.1) indeed shows not only the magnitude but also the correct direction of the force $\mathbf{F}_{e12}$. Does Eq. (3.1) need an additional explanation in words? Explain.

**Q3.2.** Would the vector form of Coulomb's law (Eq. 3.1) be possible if plus and minus signs were not associated with the two types of charges? For example, suppose that they were denoted by subscripts $A$ and $B$ instead of plus and minus signs. Explain your answer. (This question is intended to show how important proper conventions are for simplifying the mathematical description of physical phenomena.)

**Q3.3.** Is it possible to *derive* the principle of superposition of Coulomb's forces, starting from Coulomb's law? Explain.

**Q3.4.** Prove that there can be no net electric force on an isolated charged body due to its charge only.

**Q3.5.** Similarly to the electric field, the gravitational field also acts "at a distance." But whereas we understand and accept that there is a downward force on an object we lift (e.g., a stone), with no visible reason, such an electric force with no visible reason is somewhat astonishing. Explain why this is so.

**Q3.6.** Of five equal conducting balls one is charged with a charge $Q$, and the other four are not charged. Find all possible charges the balls can obtain by touching one another, assuming that two balls are allowed to touch only once, and that while two balls are touching, the influence of the other three can be neglected.

**Q3.7.** If an electrified body (e.g., a plastic ruler rubbed against a wool cloth) is brought near small pieces of thin aluminum foil, you will see that the body first attracts, but after the contact repels, the small pieces. Perform this experiment and explain.

**Q3.8.** Imagine that you electrified a body, e.g., by rubbing it against another body. How could you determine the sign of the charge on the body? Try to perform the experiment.

**Q3.9.** You have two identical small metal balls. How can you obtain identical charges on them?

**Q3.10.** Two small balls carry charges of unknown signs and magnitudes. Experiment shows that there is no electric force on a third charged ball placed at the midpoint between the first two. What can you conclude about the charges on the first two balls?

**Q3.11.** An uncharged small ball is introduced into the electric field of a point charge. Is there a force on the ball? Explain.

**Q3.12.** Is it correct to write the following: (1) $Q$ ($Q > 0$); (2) $-Q$ ($Q < 0$); (3) $Q$ ($Q < 0$); and (4) $-Q$ ($Q > 0$)? Explain.

**Q3.13.** To measure the electric field strength at a distance $r$ from a small charge $Q$, a test charge $\Delta Q$ ($\Delta Q \ll Q$) in the form of a sphere of radius $a = r/2$ is centered at that point. Discuss the correctness of the measurement.

**Q3.14.** What would the form of the expression in Eq. (3.4) be if $\mathbf{u}_r$ is *toward* the charge? What form does Eq. (3.4) take if we do not associate a sign with the charge $Q$?

**Q3.15.** Is $\rho$ in Eq. (3.6) a function of coordinates, in general?

**Q3.16.** Assuming $\rho$ in Eq. (3.8) to be known, explain in detail how you would numerically evaluate the *vector* integral to obtain **E**.

**Q3.17.** Repeat question Q3.16 for a surface distribution of charges over a surface $S$, and for a line distribution of charges along a line $L$.

**Q3.18.** Why are the formulas in Eqs. (3.8), (3.11), and (3.14) only of limited practical value?

## PROBLEMS

**P3.1.** What would be the charge of a copper cube, 1 cm on a side, if one electron were removed from all the atoms on the cube surface? A cubic meter of copper has about $8.4 \cdot 10^{28}$ atoms.

**P3.2.** Evaluate the force that would exist between two cubes as described in problem P3.1 when they are (1) $d = 1\,\text{m}$, and (2) $d = 1\,\text{km}$ apart.

**P3.3.** Three small charged bodies arranged along a straight line are at distances $a$, $b$, and $(a + b)$ apart. Determine the conditions that the charges on the bodies have to satisfy so that the electric forces on all three are zero.

**P3.4.** Assume that the earth is electrified by a charge $2Q$, and the moon by a charge $Q$. How large does $Q$ have to be so that the repulsive electric force between the earth and the moon becomes equal to the attractive gravitational force? The masses of the earth and the moon are $m_e = 5.983 \cdot 10^{24}\,\text{kg}$ and $m_m = 7.347 \cdot 10^{22}\,\text{kg}$. The gravitational constant is $\gamma = 6.67 \cdot 10^{-11}\,\text{N} \cdot \text{m}^2/\text{kg}^2$.

**P3.5.** Evaluate the specific charge of the electron (charge over mass). Estimate the charge of the book you are reading if it had the same specific charge. What would the force between two such charged books be if they were at a distance of 10 m?

**P3.6.** A given charge $Q$ is divided between two small bodies, so that one has the charge $Q'$, and the other has the rest. Determine the ratio $Q/Q'$ resulting in the greatest electric force between them, assuming the distance between them is fixed.

*Both charges have the same sign.*

**P3.7.** Three small charged bodies of charge $Q$ are placed at three vertices of an equilateral triangle with sides of length $a$. What is the direction and magnitude of the electric force on each of them if $a = 3\,\mathrm{cm}$ and $Q = 1.8 \cdot 10^{-10}\,\mathrm{C}$?

**P3.8.** A charge $Q$ exists at all vertices of a cube with sides of length $a$. Determine the direction and magnitude of the electric force on one of the charges.

**P3.9.** Two identical small, conducting balls with centers that are $d$ apart have charges $Q_1$ and $Q_2$. The balls are brought into contact and then returned to their original positions. Determine the electric force if charges $Q_1$ and $Q_2$ are (1) of the same sign; (2) of opposite signs.

**P3.10.** Evaluate the velocity of an electron orbiting around the nucleus of a hydrogen atom along an approximately circular orbit of radius $a = 0.528 \cdot 10^{-10}\,\mathrm{m}$. How many revolutions does the electron make in one second?

**P3.11.** Two small balls of mass $m$ each have a charge $Q$ and are suspended at a common point by separate thin, light, conducting filaments of length $l$. Assuming the charges are located approximately at the centers of the balls, find the angle $\alpha$ between the filaments. Suppose that $\alpha$ is small. (Such a system can be used as a primitive device for measuring charge, and is called an *electroscope*.)

**P3.12.** A small body with a charge $Q = 1.8 \cdot 10^{-10}\,\mathrm{C}$ is situated at a point $A$ in the electric field. The electric force on the body has an intensity $F = 5.4 \cdot 10^{-4}\,\mathrm{N}$. Evaluate the magnitude of the electric field strength vector at that point.

**P3.13.** A point charge $Q$ ($Q > 0$) is located at the point $(0, d/2)$, and a charge $-Q$ at the point $(0, -d/2)$, of a rectangular coordinate system. Determine and plot the magnitude of the total electric field strength vector at any point in the $xy$ plane.

**P3.14.** An electric dipole consists of two equal and opposite point charges $Q$ and $-Q$ that are a distance $d$ apart, Fig. P3.14. (1) Find the electric field vector along the $x$ axis in the figure. (2) Find the electric field vector along the $y$ axis. (3) How does the electric field strength behave at distances $x \gg d$ and $y \gg d$ away from the dipole? How does this behavior compare to that of the field of a single point charge?
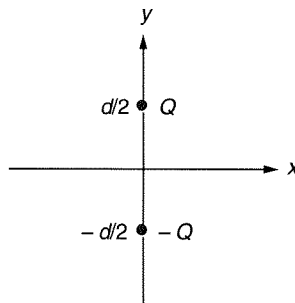


Figure P3.14 An electric dipole consists of two equal charges of opposite signs.

**P3.15.** Find the $x$ and $y$ components of the electric field vector at an arbitrary point in the field of the electric dipole from problem P3.14, assuming that the distance of the observation point from the dipole center is much greater than $d$. Plot your results.

**P3.16.** A thin, straight rod $a = 10\,\text{cm}$ long is uniformly charged along its length with a total charge $Q = 2 \cdot 10^{-10}\,\text{C}$. The rod extends from the point $(-a/2, 0)$ to the point $(a/2, 0)$ in an $xy$ rectangular coordinate system. Evaluate the electric field strength vector at points $A(0, a/4)$ and $B(3a/4, 0)$.

**P3.17.** Solve problem P3.16 approximately, by dividing the rod into $n$ segments. Compare the results with the exact solution for $n = 1, 2, 3, 4, 5, 6, 10$, and $20$.

**\*P3.18.** An L-shaped rod with sides $a = 10\,\text{cm}$ extends from the origin of an $xy$ rectangular system to the point $(a, 0)$, and from the origin to the point $(0, a)$. The rod is charged uniformly along its length with a total charge $Q = 2.6 \cdot 10^{-9}\,\text{C}$. Evaluate the electric field strength vector at points $A(a, a)$ and $B(3a/2, 0)$.

**\*P3.19.** Solve problem P3.18 approximately, by dividing the L-shaped rod into $2n$ segments. Compare the results with the exact solution for $n = 2, 3, 4, 5, 6$, and $20$.

**P3.20.** A thin ring of radius $a$ is uniformly charged along its length with a total charge $Q$. Determine the electric field strength along the ring axis.

**P3.21.** A thin circular disk of radius $a$ is charged uniformly over its surface with a total charge $Q$. Determine the electric field strength along the disk axis normal to its plane and plot your result. What do you expect the expression for the electric field to become at large distances from the disk? What do you expect the expression to become if the radius of the disk increases indefinitely, and the surface charge density is kept constant?

**P3.22.** Calculate the electric field along the axis of the disk in problem P3.21 if the charge is not distributed uniformly but increases linearly along the disk radius, and it is zero at the disk center. Plot your result and compare it to those for problem P3.21.

**P3.23.** A dielectric cube with sides of length $a$ is charged over its volume with a charge density $\rho(x) = \rho_0 x/a$, where $x$ is the normal distance from one side of the cube. Determine the charge of the cube.

**P3.24.** The volume charge density in a spherical charged cloud of radius $a$ is $\rho(r) = \rho_0(a-r)/a$, where $r$ is the distance from the cloud center, and $\rho_0$ is a constant. Determine the charge of the cloud.

**P3.25.** Determine and plot the electric field strength at a distance $r$ from a straight, very long, thin charged filament with a charge $Q'$ per unit length.

**P3.26.** A wire in the form of a semicircle of radius $a$ is charged with a total charge $Q$. Assuming the charge to be uniformly distributed along the wire, determine the electric field strength vector at the center of the semicircle.

**P3.27.** A hemispherical shell of radius $a$ is charged uniformly over its surface by a total charge $Q$. Determine the electric field strength at the center of the sphere, one-half of which is the shell.

# 4

# *The Electric Scalar Potential*

## 4.1   Introduction

You may recall from your physics courses that the gravitational field at any point can be described in two ways. One is by the force acting on a small mass located at that point. This is analogous to the electric force. The other is by specifying how large the *energy* at the point of that small mass is, per unit mass. This is known as the gravitational potential. The electric potential is analogous to the gravitational potential. It tells us how large the energy of a small charge at a point of the electric field is, per unit charge. Note that force is a *vector*, whereas energy is a *scalar*. Therefore the description of the field in terms of the potential is mathematically simpler than in terms of the field vector.

## 4.2   Definition of the Electric Scalar Potential

To understand the concept of the electric scalar potential, consider a test charge, $\Delta Q$, at a point $A$ in an electrostatic field. The electric force acting on $\Delta Q$ will tend to move the test charge. Let the force move the charge along a line $a$ to a point $B$, as sketched in Fig. 4.1. How much work was done by the electric force in this case?

41
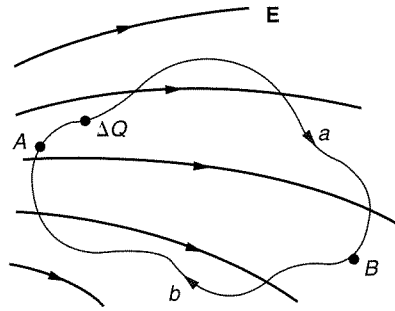
Figure 4.1 A test charge in an electrostatic field

We know from physics that if a force **F** moves a body along a small vector distance d**l**, the work done by the force is

$$dA = F\, dl \cos(\text{angle between vectors } \mathbf{F} \text{ and } d\mathbf{l}) \qquad \text{joules (J)}, \tag{4.1}$$

where $F$ and $dl$ are the magnitudes of the two vectors.

This type of product of two vectors occurs frequently in physics and engineering. (If necessary, before proceeding further please read Section A1.2 of Appendix 1.) It is known as the *scalar product*, or *dot product*. For any two vectors **X** and **Y**, the dot product is defined as

$$\mathbf{X} \cdot \mathbf{Y} = X\,Y \cos(\text{angle between vectors } \mathbf{X} \text{ and } \mathbf{Y}). \tag{4.2}$$

Hence, instead of Eq. (4.1) we can use the shorthand

$$dA = \mathbf{F} \cdot d\mathbf{l}. \tag{4.3}$$

Work is a *scalar* quantity. Therefore, to obtain the work done by the electric force in moving the test charge from point $A$ to point $B$, we simply add all elemental works of the form as in Eq. (4.3), from $A$ to $B$:

$$A_{\text{from } A \text{ to } B} = \int_A^B \mathbf{F} \cdot d\mathbf{l} \qquad \text{(J)}. \tag{4.4}$$

This type of integral (which is nothing but a sum of many very small terms) is known as a *line integral.*

The electric force on $\Delta Q$ is $\Delta Q\mathbf{E}$, and $\Delta Q$ is a constant that can be taken out of the integral sign. With this in mind, if we divide Eq. (4.4) by $\Delta Q$, we get

$$\frac{A_{\text{from } A \text{ to } B}}{\Delta Q} = \int_A^B \mathbf{E} \cdot d\mathbf{l} \qquad \text{(J/C = V)}. \tag{4.5}$$

Note that the right-hand side of this equation does *not* depend on $\Delta Q$. It represents *the work that would be done by the electric field in moving the test charge from point $A$ to point $B$*, per unit charge.

Imagine now that the field moves the test charge from $A$ to $B$ but along a different path, for example path $b$ in Fig. 4.1. How much work is done by the electric

forces in that case? It is easy to understand that the answer must be the same as for path *a*. Assume for a moment that the work that the electric field does when moving the charge along path *b* is larger than the work done along path *a*. We could then let the field move the test charge along *b* first. At point *B*, it would have a certain velocity, which means a certain kinetic energy. This energy would be greater than the work that needs to be done to return the test charge to point *A* along path *a*. So we would come back to *A* with extra energy, in spite of the system being again the same as in the beginning. Evidently, this is contrary to the law of conservation of energy. Therefore *the work done by the field or against the field in moving the test charge from one point of the field to another does not depend on the particular path between the two points.*

Since this is so, we can adopt the point *B* to be a *fixed point* and call it the *reference point, R*. We can next describe the field at all other points by specifying how large the expression in Eq. (4.5) is at these points. With the adoption of the fixed reference point $R = B$, we in fact have a *scalar* function of coordinates describing the field. It is known as the *electric scalar potential, $V_A$, at a point A of the electric field:*

$$V_A = \int_A^R \mathbf{E} \cdot d\mathbf{l} \qquad \text{volts (V).} \tag{4.6}$$

*(Definition of the electric scalar potential)*

The unit of potential is the *volt* (abbreviated V), hence the unit V/m for **E**.

We have convinced ourselves that the line integral of **E** in an electrostatic field between two points does not depend on the path of integration. An important conclusion follows from this property: the line integral of the electric field strength along *any* closed contour *C* is zero:

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = 0. \tag{4.7}$$

*(Law of conservation of energy in the electrostatic field)*

Note that the contour *C* can be completely, but also only partly, in the field. The integral on the left side of this equation is known as a *contour integral.* Note also that Eq. (4.7) represents the mathematical expression of the law of conservation of energy for the electrostatic field. It is, therefore, a fundamental property of the electrostatic field.

*Questions and problems:*   Q4.1 to Q4.8

## 4.3    Electric Scalar Potential of a Given Charge Distribution

Let us determine the potential at a point *A*, which is a distance *r* away from a positive point charge *Q*. Assume that the reference point is a distance $r_R$ away from the
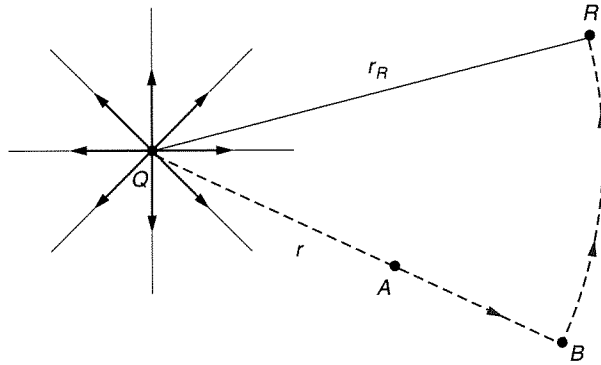
Figure 4.2 A field point, *A*, and the reference point, *R*, in the field of a point charge

charge, Fig. 4.2. We know that we can go from *A* to *R* along any path, so we adopt the simplest route: we first go from point *A* along a radius to the point *B* where it intersects with the circle of radius $r_R$. Along this path segment, vectors **E** and d**l** are parallel. Therefore the product **E** · d**l** is simply $E\,dl$ (cosine of zero is unity). We then continue to the point *R* along the arc of the circle, where the product **E** · d**l** is zero (cosine of $\pi/2$ is zero). We thus have

$$V_A = \int_A^B \mathbf{E} \cdot d\mathbf{l} + \int_B^R \mathbf{E} \cdot d\mathbf{l} = \frac{Q}{4\pi\epsilon_0} \int_r^{r_R} \frac{dr}{r^2}. \tag{4.8}$$

The integral is a standard one, and the result is

$$V_A = \frac{Q}{4\pi\epsilon_0} \left( \frac{1}{r} - \frac{1}{r_R} \right) \quad \text{(V)}. \tag{4.9}$$

This is the formula for the potential of a point charge at a distance *r* from the charge, and with respect to the reference point a distance $r_R$ from it.

So far, we have not discussed where the reference point should be. This can be *any* point. It is convenient to adopt it so that the expression for the potential is the simplest. In the case of Eq. (4.9), this is obtained if we assume that $r_R$ is very large, theoretically infinite, i.e., if the reference point is at infinity. In that case the potential at point *A* of the field of a point charge *Q* becomes

$$V_A = \frac{Q}{4\pi\epsilon_0 r} \quad \text{(reference point at infinity)} \quad \text{(V)}. \tag{4.10}$$

*(Potential at a distance r from a point charge)*

The reference point at infinity is the most convenient and is used most often. We shall see, however, that this point cannot be used if there are charges at infinity (e.g., for an infinitely long line charge).

How does the choice of the reference point influence the scalar potential function? For example, what happens if instead of $R$ we adopt $R_1$ to be the new reference point? It is left to the reader to prove that in that case the potential at all points will be increased by the *same* amount:

$$\Delta V = \int_R^{R_1} \mathbf{E} \cdot d\mathbf{l}. \tag{4.11}$$

Once we know the potential of a point charge, it is quite simple to determine the potential of a given distribution of volume, surface, or line charges. Referring to Fig. 4.3, the potential of a volume charge distribution is given by

$$V_P = \frac{1}{4\pi\epsilon_0} \int_v \frac{\rho \, dv}{r} \quad \text{(reference point at infinity)} \quad \text{(V)}, \tag{4.12a}$$

*(Potential of volume distribution of charges)*

that of a surface charge distribution is obtained as

(a)

(b)

(c)
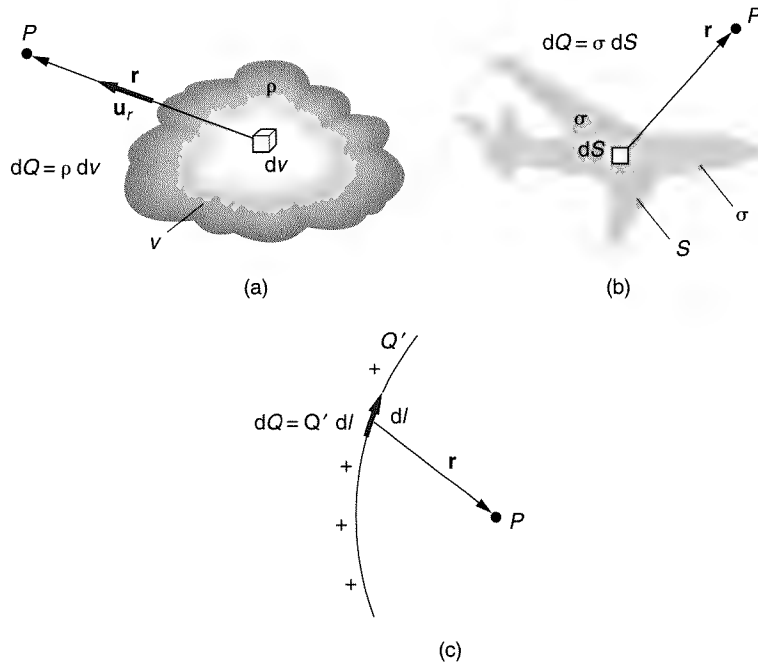
Figure 4.3 (a) A charged cloud, (b) a charged surface, and (c) a charged line, with a point $P$ at which the electric scalar potential is calculated

$$V_P = \frac{1}{4\pi\epsilon_0} \int_S \frac{\sigma\, dS}{r} \quad \text{(reference point at infinity)} \qquad \text{(V)}, \qquad (4.12b)$$

*(Potential of surface distribution of charges)*

and the potential of a line charge distribution is, by analogy,

$$V_P = \frac{1}{4\pi\epsilon_0} \int_L \frac{Q'\, dl}{r} \quad \text{(reference point at infinity)} \qquad \text{(V)}. \qquad (4.12c)$$

*(Potential of line distribution of charges)*

**Example 4.1—Potential on the axis of a charged ring.** Let us find the potential on the axis of a thin ring of radius $R$, uniformly charged along its length with a line charge density $Q'$, Fig. 4.4. The element $dl$ of the ring has a charge $dQ = [Q/(2\pi R)]\, dl$. The potential due to this charge is the same as that of a point charge, except that $Q$ needs to be replaced by $dQ$. The potential at a point $P$ on the ring axis (Fig. 4.4) is therefore obtained as

$$V_P = \frac{1}{4\pi\epsilon_0} \int_C \frac{Q}{2\pi R} \frac{dl}{r} = \frac{Q}{8\pi^2\epsilon_0 R r} \int_{\text{ring}} dl.$$

Since the integral of $dl$ around the ring equals its circumference, $2\pi R$, we finally obtain

$$V_P = \frac{Q}{4\pi\epsilon r} = \frac{Q}{4\pi\epsilon_0\sqrt{R^2 + x^2}}.$$

As already mentioned, we rarely know what the distribution of charges is. Therefore these formulas, similarly to those for the electric field strength of a given distribution of charges, do not have wide practical applicability. However, as in the case of the electric field strength, Eq. (4.12b), for example, can be used to calculate the charge distribution over a conducting body numerically.

*Questions and problems:*   Q4.9 to Q4.13, P4.1 to P4.9
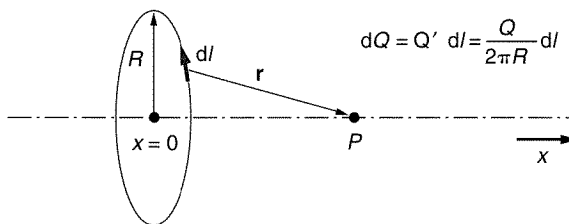


**Figure 4.4** A thin ring uniformly charged along its length with a line charge density $Q'$

## 4.4     Potential Difference and Voltage

An important concept in circuit theory is the *potential difference* or *voltage* between two points in an electrostatic field. We shall see that voltage is a wider concept than just potential difference. *Only in the electrostatic field are the two concepts equivalent.*

We denote the voltage with the same letter $V$ as the potential, either with two subscripts or with no subscripts at all (such as in the case of the potential difference between two terminals of a voltage source). The two subscripts tell us between which two points the potential difference is considered—for example, $V_{12}$ is the voltage between points 1 and 2. In the case of the potential at a point, of course, there is only one subscript, for example $V_1$, although in electrostatics we could also write it as $V_{1R}$, where $R$ denotes the reference point.

According to the definition of the potential in Eq. (4.6), the potential difference between points $A$ and $B$ is given by

$$V_{AB} = V_A - V_B = \int_A^R \mathbf{E} \cdot d\mathbf{l} - \int_B^R \mathbf{E} \cdot d\mathbf{l} \qquad \text{(V)}. \tag{4.13}$$

If in the second integral the upper and lower limits of integration are interchanged, the line element, d**l**, changes sign. Hence we can rewrite Eq. (4.13) as

$$V_{AB} = \int_A^R \mathbf{E} \cdot d\mathbf{l} + \int_R^B \mathbf{E} \cdot d\mathbf{l}. \tag{4.14}$$

So we have to integrate the dot product $\mathbf{E} \cdot d\mathbf{l}$ from $A$ to $R$, and then from $R$ to $B$, i.e., from $A$ to $B$ over $R$. We know, however, that the path between $A$ and $B$ does not affect the result. Therefore we can calculate this integral along any path, not necessarily traversing point $R$. Thus we finally have

$$V_{AB} = \int_A^B \mathbf{E} \cdot d\mathbf{l} \qquad \text{(V)}. \tag{4.15}$$

*(Potential difference between points A and B)*

Consequently, the position of the reference point does not influence the voltage between two points in an electrostatic field. This, of course, was to be expected—we know that a change in the position of the reference point changes the potential at all points by the same amount, $\Delta V$ in Eq. (4.11).

If we compare Eqs. (4.15) and (4.5), we see that the potential difference between two points can be given the following physical interpretation: it equals the work that would be done by the electric forces in moving a test charge from the first to the second point, per unit test charge.

What is the range of voltages encountered in practice? The smallest (time-varying) voltage we can measure is on the order of $1\,\text{pV} = 10^{-12}\,\text{V}$. The voltage of batteries for watches and calculators is about 1.5 V. The voltage in the plugs in our

homes is, for example, 110 V in the United States and Canada, and 220 V in Europe. The largest voltage used in power transmission by high-voltage transmission lines is on the order of $1\,\text{MV} = 10^6\,\text{V}$.

*Questions and problems:*    Q4.14 to Q4.17, P4.10 to P4.14

## 4.5    Evaluation of Electric Field Strength from Potential

Here is the final basic question we may ask about the electric scalar potential $V$: we know how to determine $V$ if we know $\mathbf{E}$ along any path from $A$ to $R$, but can we determine $\mathbf{E}$ if we know $V$? This is quite simple to do.

Consider two *close* points, $A$ and $B$, in an electrostatic field (Fig. 4.5). Let the potential at $A$ be $V_A$, and at $B$ be $V_B = V_A + dV$. Assume that the vector line element from $A$ to $B$ is $d\mathbf{l}$, and let it be along the $x$ coordinate axis so that $dl = dx$. The potential difference between $A$ and $B$ is then simply $\mathbf{E} \cdot d\mathbf{l} = E dx \cos\alpha = E_x dx$ [the integral in Eq. (4.15) consists of a single small term]. So we have

$$V_A - V_B = V_A - (V_A + dV) = -dV = E_x\,dx. \tag{4.16}$$

In other words, the component $E_x$ of vector $\mathbf{E}$ in the $x$ direction is obtained by

$$E_x = -\frac{dV}{dx} \qquad (\text{V/m}). \tag{4.17}$$

This is a very simple result. Assume that at a point in the electrostatic field we know $V$ as a function of coordinate $x$ along an $x$ axis in any direction at that point. We can then determine the projection $E_x$ of the vector $\mathbf{E}$ on the $x$ axis at that point simply as the negative derivative of $V(x)$. The reference direction for the projection is the $x$ axis.

**Example 4.2—Electric field of a point charge found from the potential.** Consider a point charge $Q$. Let the $x$ axis be any radial line beginning at the charge. The potential is then
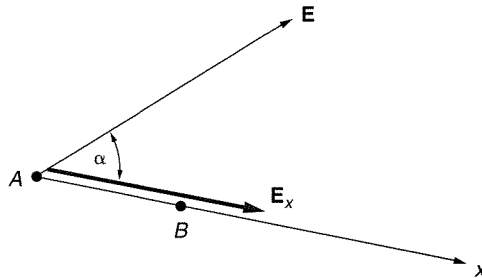


Figure 4.5 Determination of vector $\mathbf{E}$ from known $V$ at two close points

given by Eq. (4.10), except that we have to replace $r$ by $x$. According to Eq. (4.17), we have

$$E_x = -\frac{d}{dx}\left(\frac{Q}{4\pi\epsilon_0 x}\right) = \frac{Q}{4\pi\epsilon_0 x^2},$$ (4.18)

as we know it should be from Coulomb's law.

Since we now know how to determine the projection of the vector $\mathbf{E}$ in *any* direction at a point, we can easily determine the complete vector $\mathbf{E}$ at that point. We simply define three coordinate axes at the point, calculate the projections of the vector $\mathbf{E}$ on all the three axes, and sum the three components as vectors. For example, let the three axes be the $x$, $y$, and $z$ axes of a rectangular coordinate system. Then the vector $\mathbf{E}$ at any point is given by

$$\mathbf{E} = -\left(\frac{\partial V}{\partial x}\mathbf{u}_x + \frac{\partial V}{\partial y}\mathbf{u}_y + \frac{\partial V}{\partial z}\mathbf{u}_z\right) \quad \text{(V/m)},$$ (4.19)

where $\mathbf{u}_x$, $\mathbf{u}_y$ and $\mathbf{u}_z$ are unit vectors of the three coordinate axes. Partial derivatives must be used instead of ordinary derivatives because the potential $V = V(x, y, z)$ is a function of all three coordinates. To determine a projection of $\mathbf{E}$ on any one of the three coordinate axes, we have to differentiate $V(x, y, z)$ with respect to that coordinate only, considering the other two as constants. This is exactly the definition of the partial derivative of a function of several variables.

We know from mathematics that the expression in the parentheses on the right-hand side of Eq. (4.19) is called the *gradient* of the scalar function $V$. (If necessary, please read Section A1.4.1 of Appendix 1 before proceeding further.) It is sometimes written as grad $V$, but much more frequently we use the so-called *nabla operator* or *del operator*. The del operator in the rectangular coordinate system is defined as

$$\nabla = \left(\frac{\partial}{\partial x}\mathbf{u}_x + \frac{\partial}{\partial y}\mathbf{u}_y + \frac{\partial}{\partial z}\mathbf{u}_z\right) \quad \text{(1/m)},$$ (4.20)

*(Definition of nabla or del operator)*

with the assumption that the expression $\nabla V$ is a shorthand for the expression in parentheses on the right-hand side in Eq. (4.19). So we can write

$$\mathbf{E} = -\text{grad}V = -\nabla V \quad \text{(V/m)},$$ (4.21)

*(Evaluation of the electric field strength from potential)*

where, in the rectangular coordinate system,

$$\nabla V = \frac{\partial V}{\partial x}\mathbf{u}_x + \frac{\partial V}{\partial y}\mathbf{u}_y + \frac{\partial V}{\partial z}\mathbf{u}_z \qquad \underline{(V).} \qquad\qquad (4.22)$$

*(Gradient of a scalar function V in rectangular coordinates)*

**Example 4.3—Vector E on the axis of a charged ring.** As an example of the determination of **E** from the scalar potential, consider again the ring in Fig. 4.4. **E** is obtained as $-\nabla V$. The scalar potential along the ring axis is given at the end of Example 4.1. Note that it is a function of the coordinate $x$ only. Therefore at a point $x$ on the ring axis

$$\mathbf{E} = -\nabla \frac{Q}{4\pi\epsilon_0\sqrt{R^2+x^2}} = -\frac{\partial}{\partial x}\left(\frac{Q}{4\pi\epsilon_0\sqrt{R^2+x^2}}\right)\mathbf{u}_x = \frac{Qx}{4\pi\epsilon_0(R^2+x^2)^{3/2}}\mathbf{u}_x.$$

*Questions and problems:*    Q4.18 to Q4.23, P4.15 to P4.18

## 4.6    Equipotential Surfaces

A surface in an electrostatic field having the same potential at all points is called an *equipotential surface*. This is an important concept. For example, we will see that in electrostatics, the surface of any conductor is always equipotential. It can also aid in visualizing the electric field, usually in combination with electric field lines.

Since all points of an equipotential surface are at the same potential, the potential difference between two close points $A$ and $B$ on the surface is zero. Let d**l** be the position vector of point $B$ with respect to point $A$. Because d**l** is very small, the potential difference in Eq. (4.15) becomes simply $dV = \mathbf{E}\cdot d\mathbf{l}$. Since this potential difference $dV$ is zero (we assumed $A$ and $B$ to be on the same equipotential surface), *the electric field strength vector at any equipotential surface is normal to that surface.*

**Example 4.4—Equipotential surfaces in the field of a point charge.** As an example, we know that the expression for the potential of a point charge is $V(r) = Q/(4\pi\epsilon_0 r)$. Therefore the equation of the equipotential surface at a potential $V_0$ is obtained from

$$V(r) = \frac{Q}{4\pi\epsilon_0 r} = V_0,$$

from which we obtain

$$r = \frac{Q}{4\pi\epsilon_0 V_0}.$$

For different $V_0$, equipotential surfaces are spheres centered at the charge, and vector **E** is normal to these spheres.

If plotted, equipotential surfaces usually have the same potential difference from one surface to the next. Let this potential difference be $\Delta V$. For $V_0 = 0$ in the

$$r_2 = \frac{r_1}{2}$$

$$r_3 = \frac{r_1}{3} \cdots$$
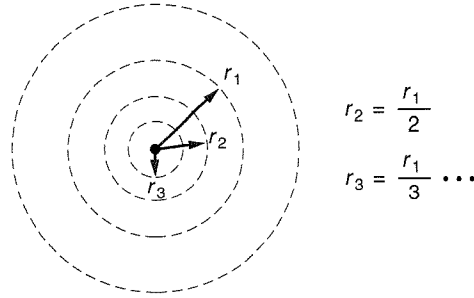
**Figure 4.6** Equipotential surfaces of the field of a point charge

preceding equation we would then have $r_0 = \infty$, for $V_0 = 1 \times \Delta V$ the radius of the equipotential surface is $r_1 = Q/(4\pi\epsilon_0 V_0)$, and so on. With this convention, therefore, equipotential surfaces for a point charge are as in Fig. 4.6.

*Questions and problems:*   Q4.24

## 4.7   Chapter Summary

1. The electric scalar potential is a scalar quantity that can be used instead of vector **E** for the description of the electrostatic field. It is defined as the line integral of **E** from any point of the field to an *arbitrary* reference point.

2. The electric scalar potential is not unique (it depends on the choice of the reference point), but for two reference points the potential at all points differs only by a constant. If there are no charges at infinity, the reference point is always adopted at infinity, but if the distribution of charges extends (theoretically) to infinity, this is not possible.

3. If we know the electric scalar potential as a function of coordinates, it is easy to obtain the component of **E** in any direction, and hence to obtain the complete vector **E**. For this, we need the mathematical concept of the gradient of a scalar function. In the rectangular coordinate system, the gradient of $V$ is obtained by the $\nabla$ operator acting on $V$, and $\mathbf{E} = -\nabla V$.

4. Being a scalar quantity, the electric scalar potential is more convenient than the vector **E** for the analysis of electrostatic fields.

5. An equipotential surface is defined as a geometrical surface with all points at the same potential. Lines of the electric field strength vector are normal to equipotential surfaces.

## QUESTIONS

**Q4.1.** Consider a uniform electric field of electric field strength $E$, and two planes normal to vector **E**, that are a distance $d$ apart. What is the work done by the field in moving a

test charge $\Delta Q$ from one plane to another? Can the work be negative? Does it depend on the location of the two points on the planes? Explain.

**Q4.2.** Is it possible to have an electrostatic field with circular closed field lines, with the vector **E** in the same direction along the entire lines? Explain.

**Q4.3.** Is it possible to have an electrostatic field with parallel lines, but of different magnitude of vector **E** in the direction normal to the lines? Explain.

**Q4.4.** If the potential of the earth were taken to be 100,000 V (instead of the usual 0 V), would it be dangerous to walk around? What influence would this have on the potential at various points, and on the difference of the potential at two points?

**Q4.5.** If we know $\mathbf{E}(x, y, z)$, is the electric scalar potential $V(x, y, z)$ determined uniquely? Explain.

**Q4.6.** Equation (4.7) is satisfied by the electric field of a point charge. Does the expression for the electric field of a point charge *follow* from Eq. (4.7)?

**Q4.7.** Why does Eq. (4.7) represent the law of conservation of energy in the electrostatic field?

**Q4.8.** What is the potential of the reference point?

**Q4.9.** As we approach a point charge $Q$ ($Q > 0$), the potential tends to infinity. Explain.

**Q4.10.** How much energy do you transfer to the electric field of a point charge when you move the reference point from a point at a distance $r_R$ from the charge to a point at infinity?

**Q4.11.** Why do we usually adopt the reference point at infinity?

**Q4.12.** Is the potential of a positively charged body always positive, and that of a negatively charged body always negative? Give examples that illustrate your conclusions.

**Q4.13.** Why are the expressions for the potential in Eqs. (4.12*a–c*) valid for a reference point at infinity?

**Q4.14.** Does it make sense to speak about voltage between a point in the field and the reference point? If it does, what is this voltage?

**Q4.15.** A charge $\Delta Q$ is moved from a point where the potential is $V_1$ to a point where the potential is $V_2$. What is the work done by the electric forces? What is the work done by the forces acting against the electric forces?

**Q4.16.** A charge $\Delta Q$ ($\Delta Q < 0$) is moved from a point at potential $V_1$ to a point at potential $V_2$. What is the work done by the electric forces?

**Q4.17.** Is $V_{AB} = -V_{BA}$? Explain.

**Q4.18.** Why is the vector **E** at a point directed toward the adjacent equipotential surface of *lower* potential?

**Q4.19.** Why do we have $\mathbf{E} = -\mathrm{grad}\, V$, and not $\mathbf{E} = +\mathrm{grad}\, V$?

**Q4.20.** A cloud of positive and negative ions is situated in an electrostatic field. Which ions will tend to move toward the points of higher potential, and which toward the points of lower potential?

**Q4.21.** Suppose that $V = 0$ at a point. Does it mean that $\mathbf{E} = 0$ at that point? Explain.

**Q4.22.** Assume we know **E** at a point. Is this sufficient to determine the potential $V$ at that point? Conversely, if we know $V$ at that point, can we determine **E**?

**Q4.23.** The potential in a region of space is constant. What is the magnitude and direction of the electric field strength vector in the region?

**Q4.24.** Prove that **E** is normal to equipotential surfaces.

# PROBLEMS

**P4.1.** Two point charges, $Q_1 = -3 \cdot 10^{-9}$ C and $Q_2 = 1.5 \cdot 10^{-9}$ C, are $r = 5$ cm apart. Find the potential at the point that lies on the line joining the two charges and halfway between them. Find the zero-potential point(s) lying on the straight line that joins the two charges.

**P4.2.** Two small bodies, with charges $Q$ ($Q > 0$) and $-Q$, are a distance $d$ apart. Determine the potential at all points with respect to the reference point at infinity. Is there a zero-potential equipotential surface? How much work do the electric forces do if the distance is increased to $2d$?

**P4.3.** A ring of radius $a$ is charged with a total charge $Q$. Determine the potential along its axis normal to the ring plane with reference to the ring center.

**P4.4.** A soap bubble of radius $R$ and very small wall thickness $a$ is at a potential $V$ with respect to the reference point at infinity. Determine the potential of a spherical drop obtained when the bubble explodes, assuming all the soap in the bubble is contained in the drop.

**P4.5.** A volume of a liquid conductor is sprayed into $N$ equal spherical drops. Then, by some appropriate method, each drop is given a potential $V$ with respect to the reference point at infinity. Finally, all these small drops are combined into a large spherical drop. Determine the potential of the large drop.

**P4.6.** Two small conducting spheres of radii $a$ and $b$ are connected by a very thin, flexible conductor of length $d$. The total charge of the system is $Q$. Assuming that $d$ is much larger than $a$ and $b$, determine the force $F$ that acts on the wire so as to extend it. Charges may be considered to be located on the two spheres only, and to be distributed uniformly over their surfaces. (Hint: when connected by the conducting wire, the spheres will be at the same potential—see Chapter 6.)

**P4.7.** Two small conducting balls of radii $a$ and $b$ are charged with charges $Q_a$ and $Q_b$, and are at a distance $d$ ($d \gg a, b$) apart. Suppose that the balls are connected with a thin conducting wire. What will the direction of flow of positive charges through the wire be? Discuss the question for various values of $Q_a$, $Q_b$, $a$, and $b$. (Hint: when connected by the conducting wire, the balls will be at the same potential—see Chapter 6.)

**\*P4.8.** The source of an electrostatic field is a volume charge distribution of finite charge density $\rho$, distributed in a finite region of space. Prove that the electric scalar potential has a finite value at all points, including the points inside the charge distribution.

**\*P4.9.** Prove that the electric scalar potential due to a surface charge distribution of density $\sigma$ over a surface $S$ is finite at all points, including the points of $S$.

**P4.10.** The reference point for the potential is changed from point $R$ to point $R'$. Prove that the potential of all points in an electric field changes by the voltage between $R$ and $R'$.

**P4.11.** Four small bodies with equal charges $Q = 0.5 \cdot 10^{-9}$ C are located at the vertices of a square with sides $a = 2$ cm. Determine the potential at the center of the square, and the voltage between the square center and a midpoint of a square side. What is the work of electric forces if one of the charges is moved to a very distant point?

**P4.12.** An insulating disk of radius $a = 5$ cm is charged by friction uniformly over its surface with a total charge of $Q = -10^{-8}$ C. Find the expression for the potential of the points which lie on the axis of the disk perpendicular to its surface. Plot your result. What are the numerical values for the potential at the center of the disk, and at a distance

$z = a$ from the center, measured along the axis? What is the voltage between these two points equal to?

**P4.13.** The volume charge density inside a spherical surface of radius $a$ is such that the electric field vector inside the sphere is pointing toward the center of the sphere, and varies with radial position as $E(r) = E_0 r/a$ ($E_0$ is a constant). Find the voltage between the center and the surface of the sphere.

**P4.14.** Two large parallel equipotential plates at potentials $V_1 = -10\,\text{V}$ and $V_2 = 55\,\text{V}$ are a distance $d = 2\,\text{cm}$ apart. Determine the electric field strength between the plates.

**P4.15.** Determine the potential along the line joining two small bodies carrying equal charges $Q$. Plot your result. Starting from that expression, prove that the electric field strength at the midpoint between the bodies is zero.

**P4.16.** Two small bodies with charges $Q_1 = 10^{-10}\,\text{C}$ and $Q_2 = -Q_1$ are a distance $d = 9\,\text{cm}$ apart. Determine the potential along the line joining the two charges, and from that expression determine the electric field strength along the line. Plot your results.

**P4.17.** From the expression for potential found in problem P4.3, find the electric field strength vector along the ring axis. (See problem P3.20.)

**P4.18.** From the general expression for the potential along the axis of the disk from problem P4.12, determine the electric field strength along the disk axis. (See problem P3.21.)

# 5

# *Gauss' Law*

## 5.1    Introduction

There is an important relation between the vector $\mathbf{E}$ in any electrostatic field and the static charge producing it. It is a consequence of the mathematical form of the electric field strength of a point charge, and is known as Gauss' law. Among other applications, Gauss' law enables a simple evaluation of the electric field in some simple but important cases.

To understand Gauss' law, we first need to understand an important mathematical concept, the *flux of a vector function through a surface*. The word "flux" originates from fluid mechanics and comes from the latin word "fluxus," which means "one that flows."

## 5.2    The Concept of Flux

Consider a uniform flow of a liquid of velocity $\mathbf{v}$ that is a function of coordinates but not of time. Imagine a net so fine that it does not disturb the flow of the liquid it is placed in. Let the surface of the net be $S$. We wish to determine the amount of the liquid that passes through the net (i.e., through $S$) in one second.

We can subdivide the surface $S$ into a large number of small flat surface elements $dS$, as in Fig. 5.1. Obviously, the total amount of liquid passing through the net is obtained as a sum of the small amounts passing through all of the small elements.
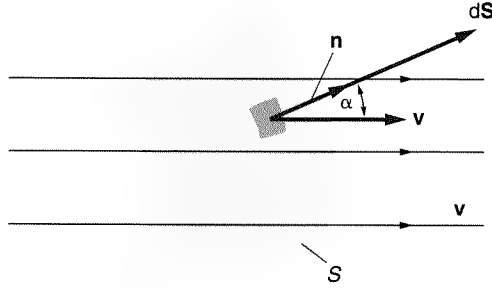
Figure 5.1 A fine net in a flow of liquid can
be approximated by a large number of flat
surface elements such as d**S**

Consider a small flat surface element shown in the figure. The vector **n** denotes a unit vector normal to the element, and $\alpha$ is the angle between this unit vector and the local velocity **v** of the fluid. It is evident that if the velocity **v** is tangential to the element, there is no flow of fluid through it. Therefore only the component of the velocity *normal* to the element contributes to the flow of liquid through the element.

In one second, the fluid at that point moves by a distance normal to d$S$ equal to $v \cos \alpha$. The quantity of fluid that passes through d$S$ in one second is therefore $v \cos \alpha$ d$S$. The quantity of fluid that passes through $S$ in one second is a sum of all these infinitely small partial flows. It is therefore an integral (an infinite sum of infinitely small terms):

$$\text{Fluid flow through } S \text{ in one second} = \int_S v \cos \alpha \, dS. \tag{5.1}$$

The expression under the integral sign has a form of a dot product, but although $v$ is the magnitude of a vector, d$S$ is not. If, however, we *define* a vector surface element d**S** as

$$d\mathbf{S} = dS \, \mathbf{n}, \tag{5.2}$$

Eq. (5.1) can be written in the form

$$\text{Fluid flow through } S \text{ in one second} = \int_S \mathbf{v} \cdot d\mathbf{S}. \tag{5.3}$$

The integral on the right side of this equation is known as the *flux of vector* **v** *through the surface S*.

It is evident that the concept of flux can be used in connection with *any* vector function, not necessarily the velocity (in which case the flux has a clear physical meaning). It is evident as well that the surface $S$ can be a closed surface. In that case, a small circle is added in the middle of the integral sign to indicate that the surface is closed.

The flux of a vector function through a closed surface is a very important concept in the theory of electromagnetic field. It is a convention to adopt the unit vector **n** normal to a closed surface *to be directed from the surface outward* (Fig. 5.2).
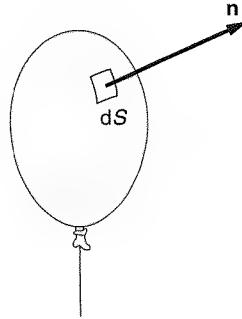
Figure 5.2 The unit vector normal to a closed surface is always adopted to be directed from the surface outward

## 5.3    Gauss' Law

Gauss' law is a very simple and important consequence of the mathematical form of the expression of the vector **E** of a point charge (i.e., of Coulomb's law). It states that the flux of the electric field strength vector through any closed surface in the electrostatic field equals the total charge enclosed by the surface, divided by $\epsilon_0$:

$$\oint_S \mathbf{E} \cdot d\mathbf{S} = \frac{Q_{\text{total in } S}}{\epsilon_0} \quad (\text{V} \cdot \text{m}). \tag{5.4}$$

*(Gauss' law)*

Basically, Gauss' law is a relationship between the sources *inside a closed surface* and the field they produce *over this entire surface*. (For interested readers, the derivation of Gauss' law is given at the end of the chapter.)

Gauss' law in Eq. (5.4) is valid for free space (air, vacuum). We know, however, that elemental charges that are actual sources of the field (electrons, protons, ions) *are* situated in a vacuum. Using this fact, we are able to extend Gauss' law to electrostatic fields in the presence of conducting and dielectric materials.

**Example 5.1—Gauss' law applied to point charges.** Consider the closed surfaces $S_1$, $S_2$, and $S_3$ in Fig. 5.3. The flux of vector **E** through $S_1$ is $(Q_1 + Q_4)/\epsilon_0$, through $S_2$ is zero, and through $S_3$ is $(Q_2 + Q_3)/\epsilon_0$.

*Questions and problems:*    Q5.1 to Q5.11, P5.1 to P5.4

Figure 5.3 Three closed surfaces, $S_1$, $S_2$ and $S_3$, in the field of four point charges

## 5.4    Applications of Gauss' Law

The applications of Gauss' law are numerous. They are basically of two kinds: proofs of some general properties of the electrostatic field, and the evaluation of the vector **E** in some special cases with high degree of symmetry of charge distribution.

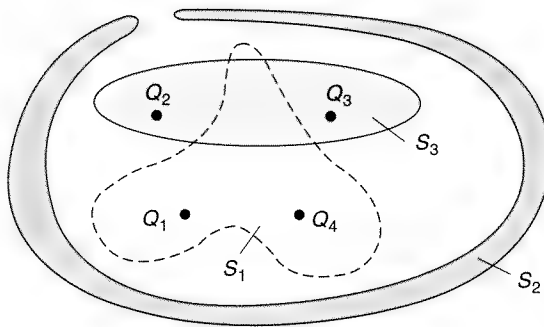**Example 5.2—Gauss' law applied to a surface of zero field.** As an example of the first kind of application, assume that we have a surface $S$ such that **E** is zero at all points of $S$. Gauss' law tells us that in *all* such cases the total enclosed charge must be zero. We will use this conclusion in the analysis of conductors in the next chapter.

Before giving further examples of Gauss' law, we note that it represents a *single* scalar equation. Therefore, in general it is not possible to determine a vector function from it (every vector function is defined by its *three* scalar components). It is possible to use Gauss' law to find **E** only if by symmetry we know everything about **E** except its magnitude.

**Example 5.3—Electric field of an infinite, charged plate.** Consider a large, theoretically infinite flat plate uniformly charged with a surface charge density $\sigma$ (Fig. 5.4a). Due to symmetry, the lines of **E** are normal to the plate, and are directed from the plate if $\sigma > 0$ and toward the plate in the other case. What we do not know is the magnitude of **E** as a function of the distance $x$ from the plate. We need one scalar equation for that, and Gauss' law can be used. Note that, from symmetry, we know that $E(-x) = E(x)$, and assume that $\sigma > 0$.

Imagine a cylinder of bases $S$ parallel to the plate and of height $2h$, positioned symmetrically with respect to the plate, as in Fig. 5.4a. Let us apply Gauss' law to that closed surface.

On the curved surface, vector **E** is parallel to it, i.e., normal to the vector surface element. Therefore, the flux of **E** through the curved surface is zero. On the two bases, vector **E** is normal to them, i.e., it is parallel to the vector surface element, so the flux of **E** through each base is simply $E(x) S$. So we have

$$\oint_{\text{cylinder}} \mathbf{E} \cdot d\mathbf{S} = E(x) S + E(-x) S = 2E(x) S = \frac{\sigma S}{\epsilon_0},$$
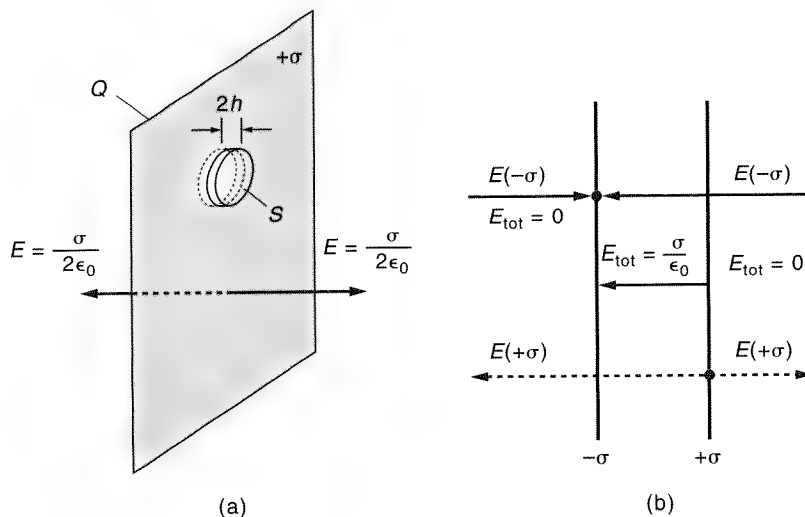
$E = \dfrac{\sigma}{2\epsilon_0}$

$E = \dfrac{\sigma}{2\epsilon_0}$

(a)

$E(-\sigma)$

$E_{tot} = 0$

$E(+\sigma)$

$E(-\sigma)$

$E_{tot} = \dfrac{\sigma}{\epsilon_0}$

$E_{tot} = 0$

$E(+\sigma)$

$-\sigma$    $+\sigma$

(b)

**Figure 5.4** (a) A charged plate, and (b) two parallel plates charged with equal surface charges of opposite sign

because the charge enclosed by $S$ is $\sigma S$. We find that the magnitude $E$ of the electric field strength *does not depend on the distance from the plate*:

$$E = \frac{\sigma}{2\epsilon_0} \quad \text{(V/m)}. \tag{5.5}$$

*(Electric field strength of uniformly charged plate)*

How is it possible that $E$ does not depend on $x$? The answer is simple. The plate being theoretically infinite, any finite distance from the plate measured with respect to the plate size is infinitely small; i.e., all points at a finite distance from the plate are equivalent.

Although we cannot have an infinite, uniformly charged plate, the result in Eq. (5.5) is nevertheless of significant importance. If we have a surface charge on a flat (or locally nearly flat) surface of any size and approach it sufficiently close and far from its edges, the field will also be given by Eq. (5.5). This is evident because from very close points the surface looks like a very large plane surface with uniform surface charge distribution (of density equal to the local surface charge density).

**Example 5.4—Electric field between two parallel charged plates.** Now consider two parallel flat plates charged with equal surface charge densities of opposite sign (Fig. 5.4b). If we have in mind the result of the preceding example, superposition yields immediately that between the plates

$$E = \frac{\sigma}{\epsilon_0} \quad \text{(V/m)}, \tag{5.6}$$

*(Electric field strength between two parallel plates with surface charges $\sigma$ and $-\sigma$)*

and that outside the plates there is no field ($\mathbf{E} = 0$). This formula may also seem unimportant for practical cases because it relates to two parallel *infinite* planes. However, this is a good approximation if the plates are of finite size but close to each other with respect to their size.

We will use Eq. (5.6) for the analysis of the parallel-plate capacitor, an important element in electrical engineering.

There are many more electrostatic systems where the magnitude of the electric field strength vector can be obtained by Gauss' law. We will consider several further important practical examples in the next chapters, when we include materials other than air (vacuum) in the analysis.

*Questions and problems:*   Q5.12, P5.5 to P5.20

## 5.5    Proof of Gauss' Law

Recall that the electric field strength vector **E** of any distribution of charge is obtained as a vector sum of individual vectors **E** resulting from all point charges of which the charge distribution is composed. Therefore Gauss' law is proven for all cases if we can prove that Eq. (5.4) is valid for a single point charge.

Consider a point charge $Q$ and let us determine the flux of vector **E** through a surface element **S** $= \mathrm{d}S\,\mathbf{n}$ (Fig. 5.5). Let us denote this flux by $\mathrm{d}\Psi_E$. It is equal to

$$\mathrm{d}\Psi_E = \frac{Q}{4\pi\epsilon_0 r^2}\mathrm{d}S\cos\alpha = \frac{Q}{4\pi\epsilon_0}\frac{\mathrm{d}S_n}{r^2}, \quad = \frac{Q}{4\pi\epsilon_0}\frac{\mathrm{d}S_n}{r^2} \tag{5.7}$$

where $\mathrm{d}S_n$ is the projection of the flat surface element $\mathrm{d}S$ on the plane normal to $r$.

The projection $\mathrm{d}S_n$ can be considered as the base of a cone with the apex at the charge. Let us cut this cone with another plane normal to $r$, for example at a distance $r_1$ from the charge, with a base of area $\mathrm{d}S_1$ (Fig. 5.5). From geometry we know that

$$\frac{\mathrm{d}S_1}{r_1^2} = \frac{\mathrm{d}S_n}{r^2}. \tag{5.8}$$

Note that $r_1$ is arbitrary. From Eq. (5.7) we conclude that the *flux through any cross-section of the cone is the same.*

Let us now enclose the charge $Q$ in Fig. 5.5 by an arbitrary closed surface $S$, indicated in the figure. We can divide this surface into elemental surfaces by a very
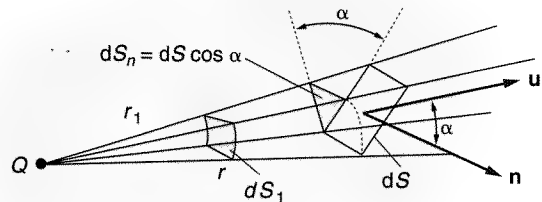


Figure 5.5 A point charge $Q$ and a surface element d**S** a distance $r$ from it

large number of cones with the common apex at the charge. To calculate the flux through any of these surfaces, we can take *any* cross-section of the cone. Therefore the flux through S is *exactly the same as that through the surface of any sphere centered at the charge.*

The flux through a sphere S of radius $r$ centered at the charge is easy to find. Noting that the angle between the vector **E** and the vector surface element d**S** of the sphere is zero and that vector **E** has the same intensity at all points on S, we have

$$\oint_S \mathbf{E} \cdot d\mathbf{S} = E \oint_S dS = E 4\pi r^2 = \frac{Q}{4\pi\epsilon_0 r^2} 4\pi r^2 = \frac{Q}{\epsilon_0}. \tag{5.9}$$

As explained, because superposition applies, this completes the proof of Gauss' law. Note that the right-hand side of Eq. (5.9) is zero if S does not enclose Q. Therefore the right-hand side in Gauss' law, Eq. (5.4), will be zero if the surface S encloses no charge.

*Questions and problems:*   Q5.13 to Q5.15

## 5.6    Chapter Summary

1. Gauss' law in Eq. (5.4) is a direct consequence of Coulomb's law, i.e., of the mathematical form of vector **E** of a point charge resulting from it. It is, therefore, a fundamental law of electrostatics.

2. In this chapter, Gauss' law has been derived for a system of charges in a vacuum. We know that the elemental charges inside matter, which are the actual sources of the electrostatic field, *are* situated in a vacuum. Therefore Gauss' law, possibly modified, should be applicable to all electrostatic fields, not only in a vacuum.

3. Gauss' law has two important types of applications. One type is proofs of certain general properties of the electrostatic field. The other is the evaluation of the intensity of vector **E** of highly symmetrical charge distributions, where we know by symmetry the direction of **E**. In such cases Gauss' law, although being a single scalar equation of one ~~scalar~~ VECTOR unknown, is sufficient to determine the magnitude of vector **E**.

.

## QUESTIONS

**Q5.1.** Prove that in a uniform electric field the flux of the electric field strength vector through any closed surface is zero.

**Q5.2.** Can the closed surface in Gauss' law be infinitesimally small in the mathematical sense? Is the answer different for the case of a vacuum and some other material? Explain.

**Q5.3.** Assume we know that the vector **E** satisfies Gauss' law in Eq. (5.4), but we do not know the expression for the vector **E** of a point charge. Can this expression be *derived* from Gauss' law? Explain.

**Q5.4.** The center of a small spherical body of radius $r$, uniformly charged over its surface with a charge $Q$, coincides with the center of one side of a cube of edge length $a$ $(a > 2r)$. What is the flux of the electric field strength vector through the cube?

**Q5.5.** A dielectric cube of edge length $a$ is charged by friction uniformly over its surface, with a surface charge density $\sigma$. What is the flux of the electric field strength vector through a slightly smaller and slightly larger imaginary cube? Do the answers look logical? Explain.

**Q5.6.** Is it possible to apply Gauss' law to a large surface enclosing a domain with a number of holes? If you think it is possible, explain how it should be done.

**Q5.7.** Inside an imaginary closed surface $S$ the total charge is zero. Does this mean that at all points of $S$ the vector **E** is zero? Explain.

**Q5.8.** A spherical rubber balloon is charged by friction uniformly over its surface. How does the electric field inside and outside the balloon change if it is periodically inflated and deflated to change its radius?

**Q5.9.** Assume that the flux of the electric field strength vector through a surface enclosing a point $A$ is the same for any size and shape of the surface. What does this tell us about the charge at $A$ or in its vicinity?

**Q5.10.** The electric field strength is zero at all points of a closed surface $S$. What is the charge enclosed by $S$?

**Q5.11.** An electric dipole (two equal charges of opposite signs) is located at the center of a sphere of radius greater than half the distance between the charges. What is the flux of vector **E** through the sphere?

**Q5.12.** Would it be possible to apply Gauss' law for the determination of the electric field for charged planes with nonuniform charge distribution? Explain.

**Q5.13.** What would be the form of Gauss' law if the unit vector normal to a closed surface were adopted to point into the surface, instead of out of the surface?

**Q5.14.** Gauss' law is a consequence of the factor $1/r^2$ in the expression for the electric field strength of a point charge (i.e., in Coulomb's law). At what step in the derivation of Gauss' law is this the condition for Gauss' law to be valid?

**Q5.15.** Try to derive Gauss' law for a hypothetical electric field where the field strength of a point charge is proportional to $1/r^k$, where $k \neq 2$.

# PROBLEMS

**P5.1.** The flux of the electric field strength vector through a closed surface is $100\,V \cdot m$. How large is the charge inside the surface?

**P5.2.** A point charge $Q = 2 \cdot 10^{-11}\,C$ is located at the center of a cube. Determine the flux of vector **E** through one side of the cube using Gauss' law.

**P5.3.** A point charge $Q = -3 \cdot 10^{-12}\,C$ is $d = 5\,cm$ away from a circular surface $S$ of radius $a = 3\,cm$ as shown in Fig. P5.3. Determine the flux of vector **E** through $S$.
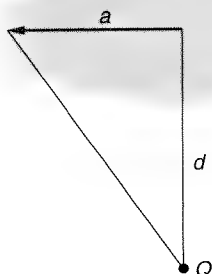
Figure P5.3 A circular surface near a point charge

**P5.4.** Determine the flux of vector **E** through a hemispherical surface of radius $a = 5$ cm, if the field is uniform, with $E = 15$ mV/m, and if vector **E** makes an angle $\alpha = 30°$ with the hemisphere axis. Use Gauss' law.

**P5.5.** Three parallel thin large charged plates have surface charge densities $-\sigma$, $2\sigma$, and $-\sigma$. Find the electric field everywhere for all combinations of the relative sheet positions and $\sigma = 10^{-6}$ C/m$^2$. Do the results depend on the distances between the plates? Determine the equipotential surfaces in all cases, and the potential difference between pairs of plates, if the distance between them is 2 cm.

**P5.6.** A very large flat plate of thickness $d$ is uniformly charged with volume charge density $\rho$. Find the electric field strength at all points. Determine the potential difference between the two boundary planes, and between the plane of symmetry of the plate and a boundary plane.

**P5.7.** The volume charge density of a thick, very large plate varies as $\rho = \rho_0 x/d$ through the plate, where $x$ is the distance from one if its boundary planes. Find the electric field strength vector everywhere. Plot your result. How large is the potential difference between the two boundary surfaces of the plate? $d$ is the thickness of the plate

**P5.8.** Two concentric spherical surfaces, of radii $a$ and $b > a$, are uniformly charged with the same amounts of charge $Q$, but of opposite signs. Find the electric field strength at all points and present your expressions graphically.

**P5.9.** The spherical surfaces from the previous problem do not have the same charge, but are charged with $Q_{inner} = 10^{-10}$ C and $Q_{outer} = -5 \cdot 10^{-11}$ C. The radii of the spheres are $a = 3$ cm and $b = 5$ cm. Find the electric field strength and potential at all points and present your expressions graphically.

**P5.10.** A spherical cloud of radius $a$ has a uniform volume charge of density $\rho = -10^{-5}$ C/m$^3$. Find the electric field strength and potential at all points and present your expressions graphically.

**P5.11.** A spherical cloud shell has a uniform volume charge of density $\rho = 10^{-3}$ C/m$^3$, an inner radius $a = 2$ cm, and an outer radius $b = 4$ cm. Find the electric field strength and potential at all points and present your expressions graphically.

**P5.12.** The volume charge density of a spherical charged cloud is not constant, but varies with the distance from the cloud center as $\rho(r) = \rho_0 r/a$. Determine the electric field strength and potential at all points. Present your results graphically.

**P5.13.** Find the expression for the electric field strength and potential between and outside two long coaxial cylinders of radii $a$ and $b$ ($b > a$), carrying charges $Q'$ and $-Q'$ per unit length. (This structure is known as a coaxial cable, or coaxial line.) Plot your results. Determine the voltage between the two cylinders.

**P5.14.** Repeat problem P5.13 assuming that the two cylinders carry unequal charges per unit length, when these charges are (1) of the same sign, and (2) of opposite signs. Plot your results and compare to problem P5.13.

**P5.15.** A very long cylindrical cloud of radius $a$ has a constant volume charge density $\rho$. Determine the electric field strength and potential at all points. Present your results graphically. Is it possible in this case to adopt the reference point at infinity? Explain.

**P5.16.** Repeat problem P5.15 assuming that the charge density is not constant, but varies with distance $r$ from the cloud axis as $\rho(r) = \rho_0 r/a$.

**P5.17.** Repeat problem P5.15 assuming that the cloud has a coaxial cavity of radius $b$ ($b < a$) with no charges.

**\*P5.18.** Prove that the electric scalar potential cannot have a maximum or a minimum value, except at points occupied by positive and negative charges, respectively.

**\*P5.19.** Prove E̲A̲R̲N̲S̲H̲A̲W̲'S̲ ~~Earnshow's~~ *theorem:* A stationary system of charges cannot be in a stable equilibrium without external nonelectric forces. (Hint: use the conclusion from problem P5.18.)

**\*P5.20.** Prove that the average potential of any sphere $S$ is equal to the potential at its center, if the charge density inside the sphere is zero at every point.

# 6

# Conductors in the Electrostatic Field

## 6.1 Introduction

Conductors are in all electric devices. They are as common in electrostatics as in other areas of electrical engineering. Nevertheless, it is important to understand how they behave in electrostatics. This behavior explains some useful electromagnetic devices. In addition, in many nonelectrostatic applications conductors behave similarly to the way they do in electrostatics. So this chapter is important beyond its application to electrostatics.

## 6.2 Behavior of Conductors in the Electrostatic Field

Conductors have a relatively large proportion of freely movable electric charges. The best conductors are metallic (silver, aluminum, copper, gold, etc.). They usually have one free electron per atom, an electron that is not bound to its atom, but moves freely in the space between atoms. Because of their small mass, these free electrons move in response to any electric field, however small, that exists inside a conductor. The

65

same is true for all other conductors, e.g., liquid solutions and semiconductors, except that inside such conductors both positive and negative free charges can exist. The number of free charge carriers is smaller and their mass greater than in metals and electrons, but this has no influence on the behavior of conductors in the electrostatic field.

Let us make an imaginary experiment. Assume that this book is a conductor. Suppose that it has both free positive and negative charges in equal number. If the book is not situated in the electric field, the number of positive and negative free charges inside any small volume is the same, and there is no surplus electric charge at any point in the book. To be more picturesque, imagine that positive charges are blue, and negative yellow. If we mix blue and yellow we get green, so your book will look green both over its surface and at any point inside.

What would happen if we establish an electric field in the book, for example, by means of two electrodes on the two sides of the book, charged with equal charges of opposite sign? Let the positive electrode be on your left. The electric field in the book will then be directed from left to right. You would notice that blue (positive) charges move from left to right (repelled by the positive electrode), and that yellow (negative) charges move from right to left. Consequently, the right side of the book will become progressively more blue (positive), and its left side progressively more yellow (negative).

The surplus charges in the body created in this manner are known as *electrostatically induced charges*. They are, of course, the source of an electric field. Because the positive induced charge is on the right side of the book, and the negative on its left side, this electric field is directed from right to left, i.e., *opposite to the initial electric field that produced the charge*. As the amount of the induced charge increases, the total field inside the book becomes progressively smaller and the motion of charges inside the book decays. In the end, the electric field of induced charges at all points inside the book cancels out the initial electric field (due to the two charged electrodes). We thus reach electrostatic equilibrium, in which there can be no electric field at any point inside our conductive book.

From this simple imaginary experiment, we conclude the following: if we have a conducting body in an electrostatic field, and wait until the drift motion of charges under the influence of the field stops (in reality, an extremely rapid process), the electric field of induced charges will *exactly* cancel out the external field, and the total electric field at all points of a conductor will be zero. Thus the first fundamental conclusion is

$$
\boxed{\text{In electrostatics, } \mathbf{E} = 0 \text{ inside conductors.} \qquad (6.1)}
$$

With this knowledge, let us apply Gauss' law to an arbitrary closed surface $S$ that is completely inside the conductor. Because vector $\mathbf{E}$ is zero at all points on $S$, the total charge enclosed by $S$ must be zero. This means that *all the excess charge (if any) must be distributed over the surfaces of conductors:*

In electrostatics, a conductor has charges only on its surface.     (6.2)

Because there is no field inside conductors, the tangential component of the electric field strength, **E**, on the very surface of conductors is also zero (otherwise it would produce organized motion of charge on its surface):

In electrostatics, $\mathbf{E}_{\text{tangential}} = 0$ on conductor surfaces.     (6.3)

Because the tangential component of **E** is zero on conductor surfaces, the potential difference between any two points of a conductor is zero. This means that the surface of a conductor in electrostatics is equipotential. Because there is no **E** inside conductors either, it follows that all points of a conductor have the same potential:

In electrostatics, the surface and volume of a conductor are equipotential.     (6.4)

Finally, a simple relation exists between the normal component, $E_n$, of **E** on a conductor surface, and the local surface charge density, $\sigma$. To derive this relation, consider a small cylindrical surface, similar to a coin, with a base $\Delta S$ and a height $\Delta h \rightarrow 0$. One base is in the conductor and the other in air (Fig. 6.1). Let us apply Gauss' law to the closed surface of the cylinder. There is no flux of **E** through the base inside the conductor (zero field) and through the infinitely narrow strip connecting the two bases (zero area). The flux of **E** through the cylinder is thus equal only to
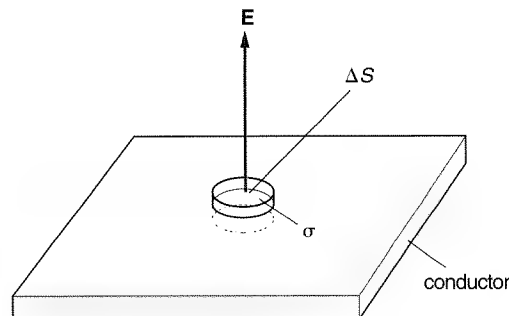


**Figure 6.1** A small cylinder of negligible height with one base in the conductor and the other in air

$E_n \Delta S$. Because the charge enclosed is $\sigma \Delta S$, using Gauss' law we obtain that on the air side of a conductor surface,

$$E_n = \frac{\sigma}{\epsilon_0} \quad \text{(V/m)}. \tag{6.5}$$

*(Normal component of electric field strength close to conductor surface)*

The simple conclusions in Eqs. (6.1) through (6.5) are all we need to know to understand the behavior of conductors in the electrostatic field.

**Example 6.1—Charged Metal Ball.** Suppose that a metal ball of radius $a$ is situated in a vacuum and has a charge $Q$. How will the charge be distributed over its surface? [We know from Eq. (6.2) that $Q$ exists only over the conductor surface.] Because equal charges repel, due to symmetry the charge distribution over the surface of the ball must be uniform. The surface charge density is therefore simply $\sigma = Q/(4\pi a^2)$. Let us determine $\mathbf{E}$ and $V$ due to this charge.

Due to the uniform charge distribution, vector $\mathbf{E}$ is radial and has the same magnitude on any spherical surface concentric to the ball. (Is such a surface an equipotential surface?) We can use Gauss' law to find the magnitude of vector $\mathbf{E}$ on any of these surfaces:

$$\oint_{\text{sphere}} \mathbf{E}(r) \cdot d\mathbf{S} = E(r) \, 4\pi r^2 = \frac{Q}{\epsilon_0}.$$

Note that the sphere encloses no charge if $r < a$. Thus

$$E(r) = \frac{Q}{4\pi \epsilon_0 r^2} = \frac{\sigma 4\pi a^2}{4\pi \epsilon_0 r^2} = \frac{\sigma a^2}{\epsilon_0 r^2} \quad (r > a), \qquad E(r) = 0 \quad (r < a). \tag{6.6}$$

This expression is the same as the one for the field of a point charge $Q$ at the center of the ball. On the surface of the ball ($r = a$), $E(a) = \sigma/\epsilon_0$, as predicted by Eq. (6.5).

It follows that outside the ball, the potential is the same as that of a point charge $Q$ placed at the center of the ball. Inside the ball the potential is constant, equal to that on the ball surface, that is,

$$V(a) = \frac{Q}{4\pi \epsilon_0 a} = \frac{\sigma a}{\epsilon_0}. \tag{6.7}$$

**Example 6.2—Charged Metal Wire.** Consider a very long (theoretically, infinitely long) straight metal wire of circular cross section of radius $a$. Let it be charged with $Q'$ per unit length. What are the field and potential everywhere around the wire?

Due to symmetry, the charge will be distributed uniformly over the wire surface. It is not difficult to conclude that, as the result of this symmetrical charge distribution, vector $\mathbf{E}$ is radial. Its magnitude depends only on the normal distance $r$ from the wire axis and can be determined by Gauss' law.

For the application of Gauss' law, we adopt the cylindrical surface shown in Fig. 6.2. There is no flux through the cylinder bases because vector $\mathbf{E}$ is tangential to them. The total flux through the closed surface is therefore equal to the flux through its cylindrical part. Applying Gauss' law gives
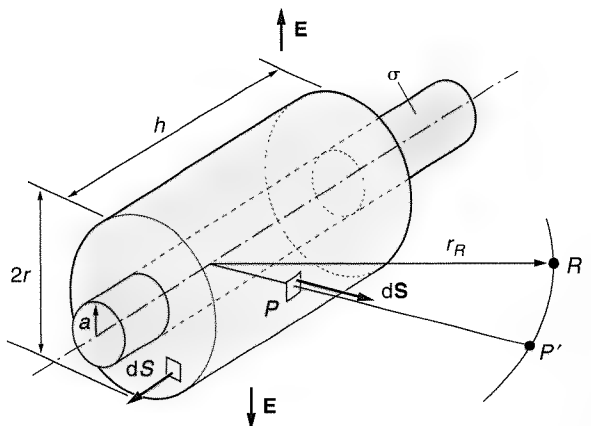
Figure 6.2 Segment of an infinitely long straight wire of circular cross section of radius $a$

$$\oint_S \mathbf{E} \cdot d\mathbf{S} = \int_{\text{belt}} \mathbf{E} \cdot d\mathbf{S} = E(r)\, 2\pi r h = \frac{Q_{\text{in cylinder}}}{\epsilon_0} = \frac{Q'h}{\epsilon_0}.$$

Note that if $r < a$ the surface encloses no charge. Thus,

$$E(r) = \frac{Q'}{2\pi \epsilon_0 r} \quad (r > a), \qquad E(r) = 0 \quad (r < a). \tag{6.8}$$

*(Electric field of straight, infinitely long, uniformly charged thin wire)*

Because the surface charge density on the cylinder is $\sigma = Q'/(2\pi a)$, $E(r)$ on the wire surface can be written in the form $E(a) = \sigma/\epsilon_0$. This, of course, is the same result as obtained by applying Eq. (6.5).

The determination of potential is slightly more complicated. Consider a point $P$ at a distance $r$ from the wire axis. Let the reference point, $R$, be at $r_R$ from the axis, in the plane containing $P$ and the wire axis. Recall that we can adopt any path from $P$ to $R$ in determining the potential. We choose the simplest: first a radial line from $P$ to the distance $r_R$ from the wire axis, and then a line parallel to the axis to $R$, Fig. 6.2. Along the first path segment, $\mathbf{E}$ and the line element are parallel, so $\mathbf{E} \cdot d\mathbf{l} = E(r)\, dl = E(r)\, dr$, because the line element, $dl$, becomes the differential increase in $r$, $dr$. Along the second path segment $\mathbf{E} \cdot d\mathbf{l} = 0$. Thus we have

$$V(r) = \int_P^R \mathbf{E} \cdot d\mathbf{l} = \int_r^{r_R} E(r)\, dr = \frac{Q'}{2\pi \epsilon_0} \int_r^{r_R} \frac{dr}{r},$$

or

$$V(r) = \frac{Q'}{2\pi \epsilon_0} \ln \frac{r_R}{r}. \tag{6.9}$$

*(Potential of straight, infinitely long, uniformly charged thin wire)*

We see that in this case we *cannot* adopt the reference point at infinity, because $\log \infty \to \infty$.

The expressions in Eqs. (6.8) and (6.9) are also useful for noninfinite wires, as long as we are interested in the field at points close to the wire and away from the ends. Because metallic wires are used often in electrical engineering, these equations are important.

*Questions and problems:*   Q6.1 to Q6.4

## 6.3    Charge Distribution on Conductive Bodies of Arbitrary Shapes

Only for symmetrical isolated conductors is the charge distribution on their surface known—actually, inferred from symmetry. For conducting bodies of arbitrary shape the determination of charge distribution is one of the most important—and the most difficult—problems in electrostatics. Except in a few relatively simple cases, it can be determined only numerically. For many applications, it is useful to have a rough idea what the charge distribution is like. In estimating the charge distribution, the following simple reasoning can be of significant help.

We know that on an isolated metal sphere the charge is distributed uniformly. We also know that if the radius of the sphere is $a$ and the surface charge density on it is $\sigma$, then the potential of the sphere is $V(a) = \sigma a/\epsilon_0$ (Eq. 6.7). Let us use this expression to estimate the charge distribution on a more complex conducting body.

Consider a charged metal body sketched in Fig. 6.3. It consists of a larger sphere of radius $a$, onto which are pressed parts of two smaller spheres of radii $b$ and $c$.

Close to points $A$, $B$, and $C$ indicated in the figure, the surface charge density is not the same. These three points are, however, at the same potential, $V$, because the body is conductive. Because charges that are close to a certain point predominantly contribute to the potential at that point, roughly speaking the surface charge density $\sigma_A$ is approximately that of a sphere of radius $a$ at the potential $V$. Therefore, according to Eq. (6.7), $\sigma_A \simeq \epsilon_0 V/a$. Similarly, $\sigma_B \simeq \epsilon_0 V/b$, and $\sigma_C \simeq \epsilon_0 V/c$. Thus, for the conducting body shown in Fig. 6.3,

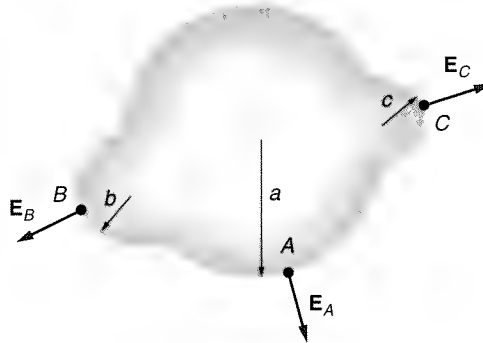$$a\sigma_A \simeq b\sigma_B \simeq c\sigma_C. \tag{6.10}$$



Figure 6.3 A charged metal body

Because the surface charge density is proportional to the local electric field strength,

$$aE_A \simeq bE_B \simeq cE_C. \tag{6.11}$$

These are simple but important approximate results. They tell us that the surface charge density at different points on a metal body is approximately inversely proportional to the curvature of the surface of the body at these points. This means that the *largest charge density and electric field strength on charged conductive bodies is around sharp parts of the body.*

An application of Eq. (6.11), for example, is a simple method for discharging aircraft. During flight, the plane becomes charged due to air friction. This charge could produce large fields during landing that in turn could produce a spark resulting in fire. However, if we place conducting spikes on the wings and other pointed plane parts, the charge density and, consequently, the electric field at these points become very high and the air ionizes (i.e., becomes conductive). A large portion of the charge "leaks" through these conducting channels into the atmosphere. We will see later that the principle behind lightning arresters is quite similar.

*Questions and problems:*    Q6.5 to Q6.8, P6.1 to P6.7

## 6.4    Electrostatic Induction

Let us reconsider the electrostatically induced charges introduced in section 6.2 from a slightly different viewpoint. Assume that the metal body *A* shown in Fig. 6.4a is charged. The charge is distributed approximately as shown. What happens if we bring an *uncharged* conductive body *B* close to body *A*, and do it very quickly (theoretically, at infinite speed)?

In slow motion, the electric field of body *A* will first move free charges inside body *B*. Assume that there are both positive and negative free charges in body *B*. The force on the positive charges will be in the direction of vector **E**, and on the negative ones in the opposite direction. Charges of opposite sign will crowd up on two sides of *B*. We know from section 6.2 that their electric field is opposite to the field of charges
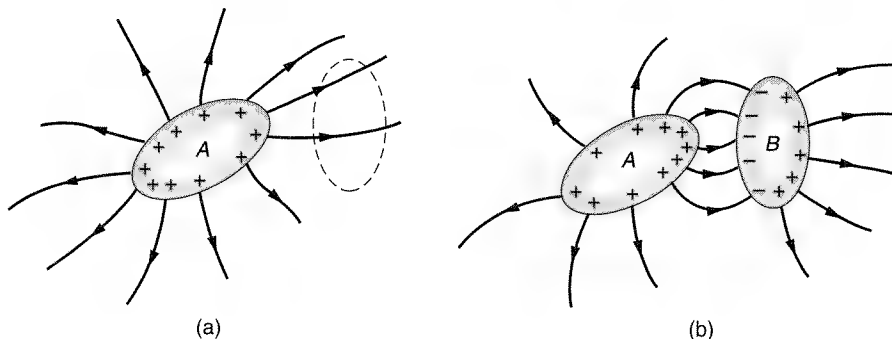


(a)                                                    (b)

**Figure 6.4** (a) A charged conducting body, and (b) approximate distribution of charge and field when an *uncharged* body, *B*, is brought near the first body, *A*

on body *A*. Once electrostatic equilibrium is reached, this field exactly cancels the field due to the charge on *A*. Of course, *the field of the charge distribution on B will change to some extent the original charge distribution on body A*. The final charge distribution and electric field lines are sketched in Fig. 6.4b. As already mentioned, the entire process of charge redistribution is very fast, practically instantaneous.

Because body *B* in the beginning was not charged, the total charge in body *B* must remain zero. However, equal charges of opposite sign do appear over the body surface. This is called *electrostatic* or *electric induction*, and we say that the charges on body *B* are *induced*. If body *B* had been charged previously, a similar process would have taken place. The charge would have redistributed itself so that the total electric field inside the body is zero.

Electrostatic induction is of great importance in electrical engineering. For example, so-called electric coupling between elements or wires in an electronic circuit (including traces on a printed-circuit board) is a result of electrostatic induction. In the examples that follow we describe some of the effects and applications of electrostatic induction.

**Example 6.3—Electrostatic induction for a conductive body in a uniform electric field.** Consider a metal sphere placed in a uniform electrostatic field, as shown in Fig. 6.5 (for example, between two very large, oppositely charged, parallel metal plates). The induced charge on the sphere will distribute itself to cancel out the uniform electric field inside it. We know that the resulting electric field on the air side is perpendicular to the surface of the sphere. The field lines will therefore "bend," as indicated in Fig. 6.5. Note that the resulting electric field is much stronger than the original at points *A* and *B*. (It can be proved that for the sphere, it is exactly three times stronger.) This is important for understanding the influence of the presence of water drops and metal particles on the electrical properties of liquid dielectrics. We will later come back to this example when we study the processes of xerography, electrostatic exhaust gas purification, and industrial electrostatic separation in Chapter 11.

**Example 6.4—Faraday's cage.** Consider an uncharged metal shell, as in Fig. 6.6. Let the shell be situated in an electrostatic field. We know that the electric field inside the conducting
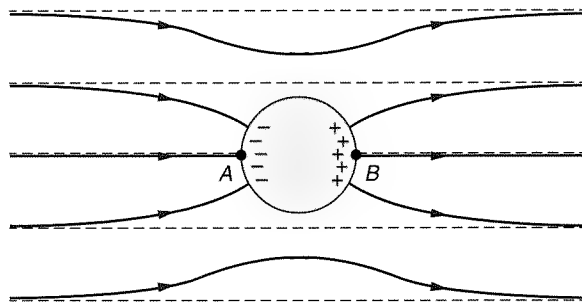


Figure 6.5 When an uncharged metal sphere is brought into a uniform electric field, the field becomes nonuniform. The strength of the field becomes significantly greater at points *A* and *B* of the sphere.
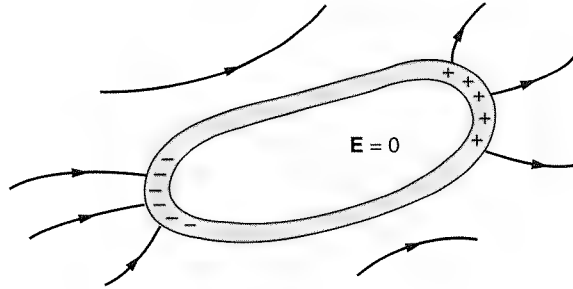
Figure 6.6 A metal shell shields its inside from electrostatic fields.

walls of the shell is zero (Eq. 6.1). It is not possible that the field penetrates through a field-free region (that would be analogous to attempting to pull something with a rope, one part of which is cut off). Therefore *there is no field inside the shell cavity either, no matter how thin the shell may be.* This means that we can shield a part of space from electrostatic fields perfectly.

It turns out that the shielding is efficient (although not any more perfect) if we use a metal grid instead of the metal wall, and that there are shielding effects even when the field is not electrostatic. Therefore such a shield can be used to protect a domain of space (for example, a small room) from external fields. Such a shielded space is known as a *Faraday cage.* It is a standard piece of equipment in many electrical engineering laboratories, both in electronics and power engineering. For example, your microwave oven door has a metal mesh, and when it is closed it forms a Faraday cage with the metal walls.

**Example 6.5—Electrostatic induction due to a point charge inside a hollow conducting sphere.** Let a point charge be inside a spherical uncharged metal shell, as in Fig. 6.7. It will induce some charge on the inside shell wall. To determine the induced charge, we apply Gauss' law on a spherical surface $S$ inside the metal wall. The field at all points of $S$ being zero, the total enclosed charge is zero as well. This means that the induced charge on the inside shell wall amounts to exactly $-Q$. Because the total shell charge is zero, and we know that inside conductors there can be no charge, a charge $Q$ appears on the outside shell wall. How is this charge distributed?

Because there is no field inside the wall, there is no connection whatsoever between the field inside the shell cavity and the field outside the shell. The outer charge, therefore, is distributed as if there were no charge inside the cavity, which means *uniformly, irrespective of the position of the charge Q inside the cavity.*

Conversely, what would happen with the field inside the cavity if we change the outer charge, remove it, or bring an electrified object near the shell? Because there is no field in the shell wall, nothing of the kind can have any effect on the field inside the cavity. Therefore we can perform any electromagnetic experiment *inside* the cavity knowing it cannot be detected from outside. This is an example of the "reverse" application of the Faraday cage.

**Example 6.6—Induced charges on the surface of the earth due to charged clouds.** Let a charged cloud be above the earth's surface, as shown in Fig. 6.8. The soil is always conductive to some extent, so charges of opposite sign to those of the cloud will be induced on the earth's surface below the cloud. (We know that somewhere far away on the planet's surface the same amount of charge, of opposite sign, will appear.)
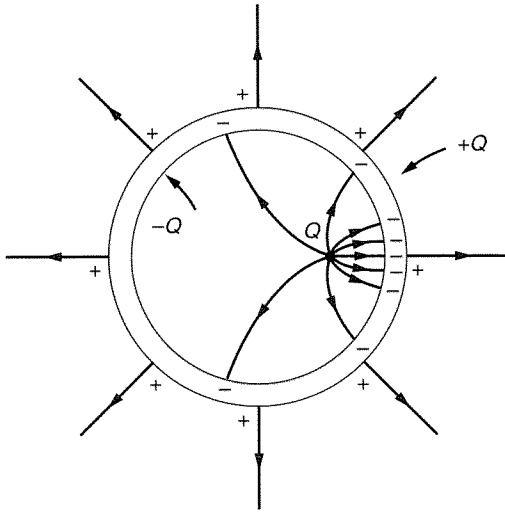
**Figure 6.7** A point charge $Q$ inside an uncharged metal spherical shell induces a total charge $-Q$ on the inside walls of the shell. The charge on the shell outside wall is distributed uniformly irrespective of the position of the point charge.
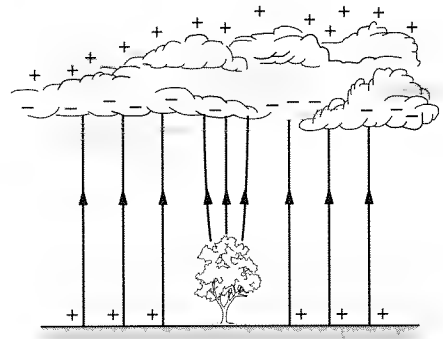


**Figure 6.8** A charged cloud above the surface of the earth

The electric field on the earth's surface below the cloud will be the strongest at higher, sharper points on it—for example, at the top of a tree or at the top of a tall building. It is frequently so strong that it provokes local ionization of air, which may extend up to the cloud in the form of lightning (the cloud discharge to the earth). To protect an object from lightning strikes, a metal spike (or a system of spikes) on the top of the roof is connected to the ground with a wire. In this way the lightning is purposely attracted to strike at a desired point, and the cloud charge is taken to the ground through the wire.

**Example 6.7—The effect of connecting a charged body to ground.** Consider a charged metallic body of charge $Q$ situated above ground. What happens if we decrease the height of the body until it touches the ground (which, as mentioned, is always conductive)? An example of such a case is shown in Fig. 6.9. An airplane gets charged by friction when flying through clouds, say with a positive charge $Q$. (The cloud remains negatively charged.) As the plane is landing, the induced charge on the ground redistributes. When the plane is at a height $h_1$ (Fig. 6.9a), there is an induced negative charge under it, spread over a large area. The remaining positive charge of the neutral earth is distributed over far areas of the earth. As the plane is landing (Fig. 6.9b,c), the induced charge becomes more and more localized below the plane. When the plane touches down, its charge $Q$ neutralizes the local induced charge and the plane is discharged. There is still leftover induced positive charge, which now redistributes itself uniformly over the entire earth, but due to the enormous size of the earth, its surface density is negligible. The overall charge of the atmosphere-earth system is still the same—the original cloud carries the negative of the remaining induced positive charge. (This charge could become part of a lightning stroke and eventually neutralize the positive induced charge.)

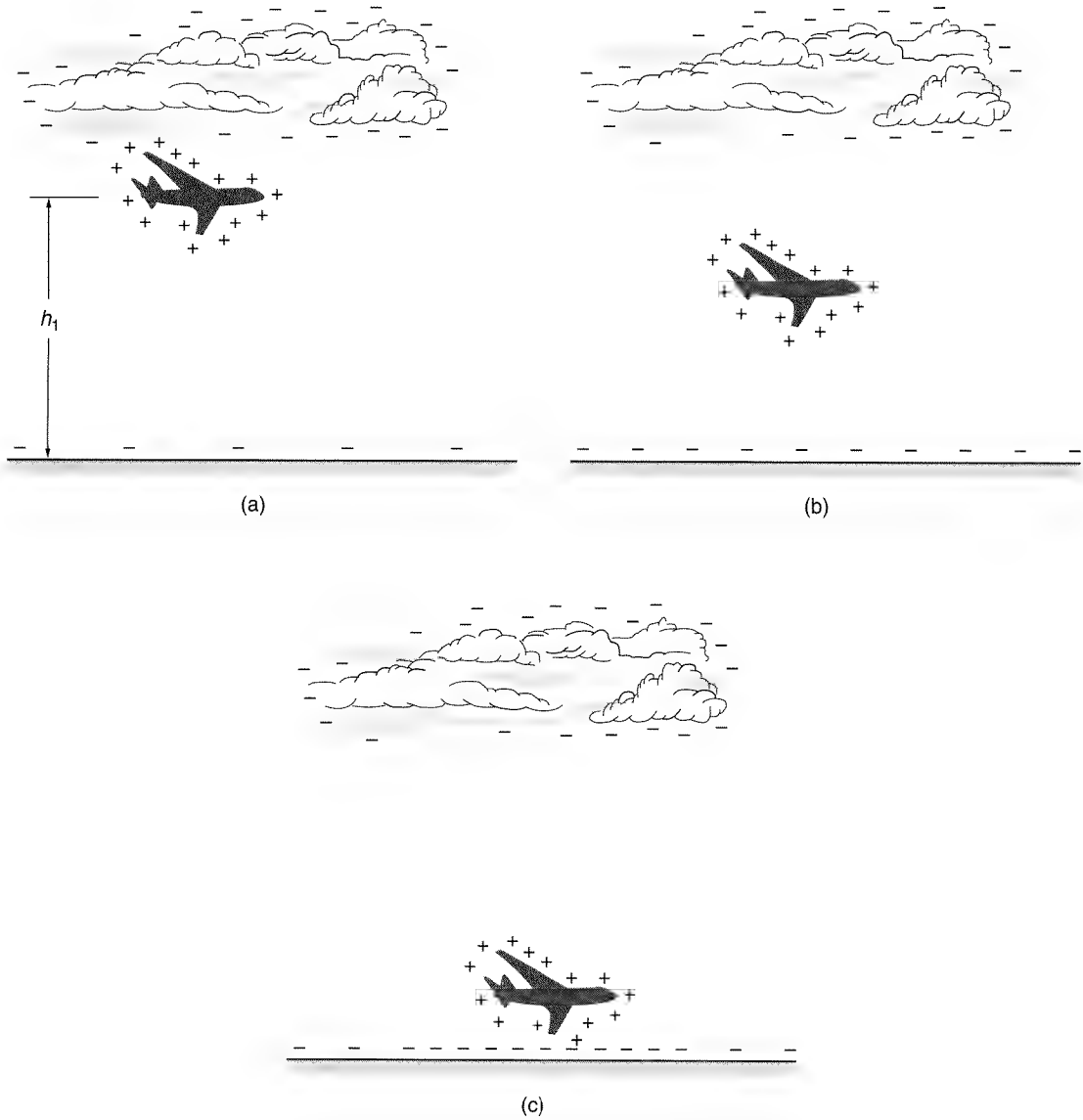*Questions and problems:*    Q6.9 to Q6.17, P6.8 to P6.12

(a)

(b)

(c)

**Figure 6.9** Induced charge when a charged airplane is landing

## 6.5    Image Method for Charges Above a Conducting Plane

The often-used *method of images* is a special case of a general theorem in electromagnetics known as the *equivalence theorem*. (Thévenin's and Norton's theorems in circuit theory are also special cases of the equivalence theorem.) The fundamental concept behind this theorem is the following.

It turns out that there are an infinite number of sources that can be placed *inside* any region of space, such that they would all produce the same field *outside* that region. For example, the field outside a spherically symmetrical cloud of radius $a$ and total charge $Q$ is the same as that due to a point charge $Q$ at its center, or to a uniform surface charge $Q$ over *any* sphere of radius less than or equal to $a$. These three sources are shown in Fig. 6.10. They are said to be equivalent with respect to the region where we are interested in the field, in this case the outside of the sphere. (Note that inside the sphere, the field is different in the three cases.) In some cases
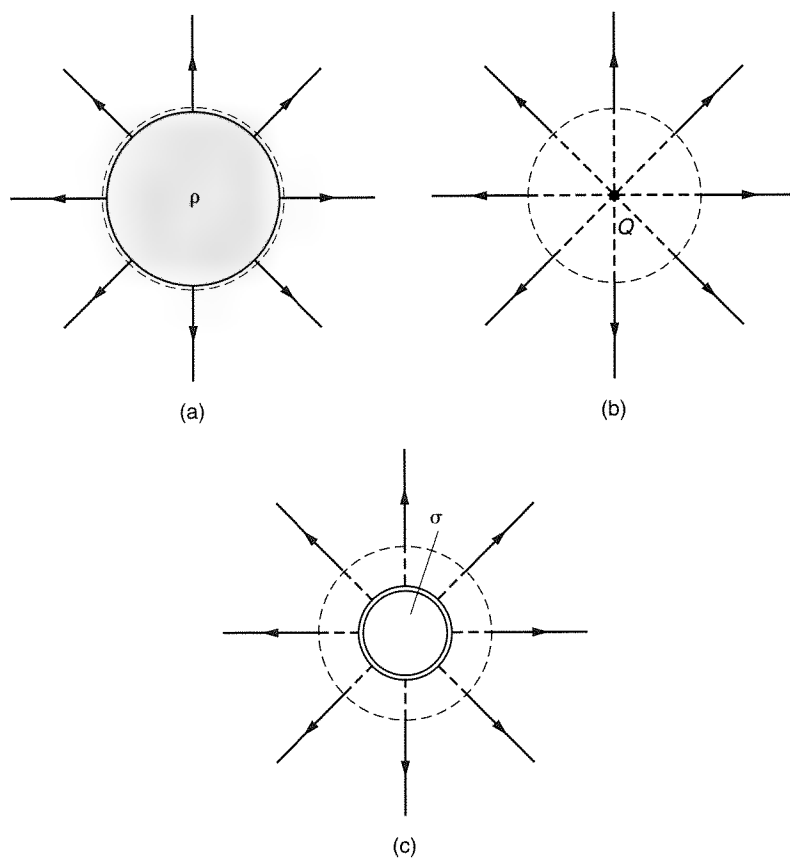


Figure 6.10 Equivalent charge distributions for the field outside a spherical surface of radius $a$: (a) charged ball of radius $a$; (b) point charge; and (c) surface-charged spherical shell of radius $r \leq a$
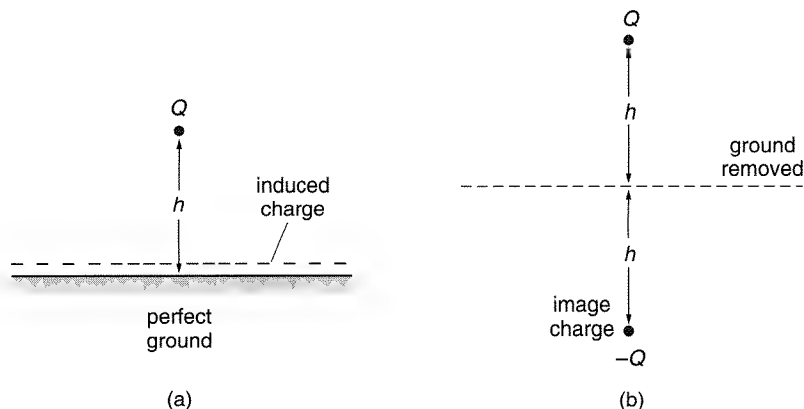
Figure 6.11 (a) Point charge above a large, grounded conducting sheet, and (b) induced charges on the sheet replaced by an equivalent point charge

it is possible to find equivalent sources that are much simpler than the actual ones. The method of images of charges in a conducting plane is one such example of great practical usefulness.

Let a point charge $Q$ be located above a very large, flat conducting sheet that is grounded, as in Fig. 6.11a. The sheet is equipotential. (For example, it can be the surface of the earth, which is usually adopted as the potential reference.) According to Gauss' law, a charge $-Q$ is induced on the upper surface of the sheet. (It is advised that the reader prove this statement as an exercise.) We know that the induced charge is distributed in such a way as to cancel the electric field inside the sheet and make the tangential component of **E** equal to zero on the surface. We do not know, however, what this distribution is like, and therefore we cannot evaluate the field it produces above the sheet.

Although it is possible to determine this distribution starting from an integral equation, there is a much simpler way of doing it. Note that two charges of the same magnitude and opposite sign result in zero tangential electric field on the plane of symmetry of the two charges. This leads us to the conclusion that a source equivalent to all these unknown induced charges, with respect to the space above the plane, is a *single point charge* $-Q$, positioned symmetrically with respect to the plane. (Of course, once the equivalent charge is in place, we remove the induced charges.) This equivalent system is sketched in Fig. 6.11b. The equivalent source $-Q$ in this case is usually referred to as the *image of the charge $Q$ in the conducting plane*. Once the ground plane is replaced by the image, the field *below* the ground plane is different than in the original system. Note that, knowing the image, we can also find the actual surface charge distribution over the metal plane (see problem P6.17).

Since superposition applies, images of any distribution of charges above a conducting plane are found in the same way. An important example is a wire at a height $h$ above ground, such as a conductor of a power line or a phone cable, with a charge $Q'$ per unit length (Fig. 6.12a). The equivalent source to the charges induced on the
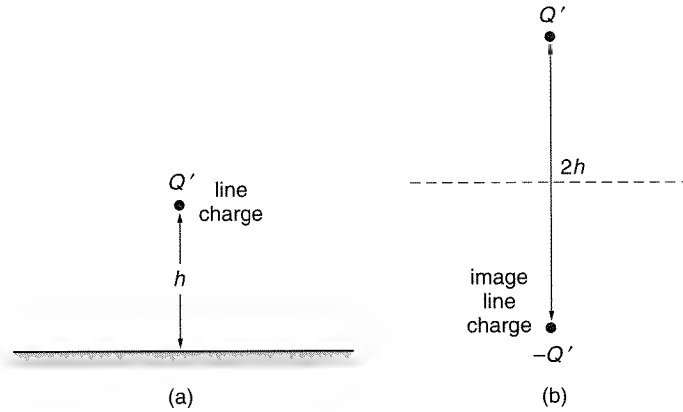
Figure 6.12 (a) Line charge above a large ground plane, and (b) induced charges on the plane replaced by an equivalent line charge

ground is simply a wire with a charge $-Q'$ per unit length situated at a depth $h$ below the former surface of the ground, with the ground removed, as in Fig. 6.12b.

*Questions and problems:*    P6.13 to P6.18

## 6.6    Chapter Summary

1. Most properties of the electrostatic field in the presence of conducting bodies can be deduced from the following:

    a. In electrostatics, the charge distribution must be such that $\mathbf{E} = 0$ inside conductors.

    b. In electrostatics, all excess charge in a conductor is spread over its surface.

    c. In electrostatics, $\mathbf{E}_{\text{tang}} = 0$ on the surface of conductors; that is, $\mathbf{E}$ is perpendicular to the surface of conductors.

    d. In electrostatics, conductor surfaces (and all points inside conductors) are equipotential.

    e. The normal component of the vector $\mathbf{E}$ on the surface of a conductor in a vacuum is equal to $\sigma/\epsilon_0$.

    Actually, the properties b to d are consequences of a.

2. With these facts, we are able to understand the phenomenon of electrostatic (or electric) induction, the physics behind electrostatic screens (Faraday cages), the functioning of lightning rods, and the physical meaning of grounding a metal object.

3. They also help us understand the concept of replacing a ground plane with an image of a charge above it. This image, situated below the original ground plane, produces the same field above the ground plane as the original induced

charges on the plane. It provides a simple way of finding the fields due to charges above a perfectly conducting ground plane.

# QUESTIONS

**Q6.1.** Prove that all points of a conducting body situated in an electrostatic field are at the same potential.

**Q6.2.** Two thin aluminum foils of area $S$ are pressed onto each other and introduced into an electrostatic field, normal to the vector $\mathbf{E}$. The foils are then separated while in the field, and moved separately out of the field. What is the charge of the foils?

**Q6.3.** An uncharged conducting body has four cavities. In every cavity there is a point charge, $-Q_1$, $-Q_2$, $-Q_3$, and $-Q_4$. What is the induced charge on the surfaces of the cavities? What is the charge over the outer surface of the body?

**Q6.4.** We know that there is no electrostatic field inside a conductor. Assume that we succeeded in producing an electric field that is tangential to a conducting body just above its surface. Is this physically possible? If you think it is not, which law do you think would be violated in that case?

**Q6.5.** If *uncharged* pieces of aluminum foil are brought close to an electrified metal body, you will notice that they will be attracted, and then some of them repelled. Explain.

**Q6.6.** If an uncharged body (e.g., your finger) is brought near a small charged body, you will notice that the body is attracted by the uncharged body (your finger). Explain.

**Q6.7.** A very thin short conducting filament is hanging from a large conducting sphere. If the sphere is charged with a charge $Q$, is the charge on the filament greater or less than that which remains on the sphere? Explain.

**Q6.8.** An uncharged conducting flat plate is brought into a uniform electrostatic field. In which position of the plate will its influence on the field distribution be minimal, and in which maximal?

**Q6.9.** Assume that the room in which you are sitting is completely covered by thin aluminum foil. To signal to a friend outside the room, you move a charge around the room. Is your friend going to receive your signal? Explain.

**Q6.10.** Assume that in question Q6.9 your friend would like to signal you by moving a charge. Would you receive his signal? Explain.

**Q6.11.** A small charged conducting body is brought to a large uncharged conducting body and connected to it. What will happen to the charge on the small body? Is this the same as if a charged conducting body is connected to the ground?

**Q6.12.** A point charge $Q$ is brought through a small hole into a thin uncharged metallic spherical shell of radius $R$, and fixed at a point that is a distance $d$ $(d < R)$ from its center. What is the electric field strength outside the shell?

**Q6.13.** A very thin metal foil is introduced exactly on a part of the equipotential surface in an electrostatic field. Is there any change in the field? Are there any induced charges on the foil surfaces? Explain.

**Q6.14.** A closed equipotential surface enclosing a total charge $Q$ is completely covered with very thin metal foil. Is there any change in the field inside and outside the foil? What is the induced charge on the inner surface of the foil, and what on the outer surface?

**Q6.15.** A thin wire segment is introduced in the field and placed so that it lies completely on an equipotential surface. Is there any change in the field? Are there any induced charges on the wire surface? Explain.

**Q6.16.** Repeat question Q6.15 assuming that the wire segment is made to follow a part of the line of vector **E**.

**Q6.17.** Describe what happens as an airplane, charged negatively by friction with a charge $-Q$, is landing and finally touches down.

# PROBLEMS

**P6.1.** A small conducting sphere of radius $a = 0.5\,\text{cm}$ is charged with a charge $Q = 2.3 \cdot 10^{-10}\,\text{C}$, and is at a distance $d = 10\,\text{m}$ from a large uncharged conducting sphere of radius $b = 0.5\,\text{m}$. The small sphere is then brought into contact with the large sphere, and moved back into its original position. Determine approximately the charges and potentials of the small and the large spheres in the final state. Take into account that $a \ll b$.

**P6.2.** A large charged conducting sphere of radius $a = 0.4\,\text{m}$ is charged with a charge $Q = -10^{-9}\,\text{C}$. A small uncharged conducting sphere of radius $b = 1\,\text{cm}$ is brought into contact with the large sphere, and then taken to a very distant point. Determine approximate charges and potentials of the large and small spheres in the end state, as well as the potential of the large sphere in the beginning.

**P6.3.** Two conducting spheres of equal radii $a = 2\,\text{cm}$ are far away from each other, and carry charges $Q_1 = -4 \cdot 10^{-9}\,\text{C}$ and $Q_2 = 2 \cdot 10^{-9}\,\text{C}$. The spheres are brought to each other, touched, and moved back to their positions. Determine the charges of the spheres in the final state, as well as the potentials of the spheres in the initial and final states.

**P6.4.** The electric field strength at a point $A$ on the surface of a very thin charged conducting shell is $E$. Determine the electric field strength in the middle of a small round hole made in the shell and centered at point $A$.

**P6.5.** Inside a spherical conducting shell of radius $b$ is a conducting sphere of radius $a$ $(a < b)$, charged with a charge $Q_a$. What is the potential*of the shell: (1) if it is uncharged? (2) if it is charged with a charge $Q_b$? Does the potential depend on the position of the sphere inside the shell? Will it change if we move the sphere into contact with the inner surface of the shell?

**P6.6.** Suppose that the shell in problem P6.5 is connected by a thin conducting wire to the reference point of the potential. Determine its charge, and determine the electrostatic potential function outside the shell.

**P6.7.** A conducting sphere of radius $a$ carries a charge $Q_1$. Concentric with the sphere is a spherical shell of inner radius $b$ $(b > a)$ and outer radius $c$, carrying a charge $Q_2$. Determine the electric field intensity and the electric scalar potential at every point of the system. Plot the dependence of $E$ and $V$ on the distance $r$ from the common center.

**P6.8.** Twenty small charged bodies each carrying a charge $Q = 10^{-10}\,\text{C}$ are brought into an uncharged metallic shell of radius $R = 5\,\text{cm}$. Evaluate the potential of the shell and the electric field strength on its surface.

**P6.9.** How large an electric charge must be brought into the shell from problem P6.8 to achieve a field of $30\,\text{kV/cm}$ at its surface? (This is approximately the greatest electric

field strength in air; for larger fields, the air ionizes and becomes a conductor, or breaks down.)

**P6.10.** A metal shell with a small hole is connected to ground with a conducting wire. A small charged body with a charge $Q$ ($Q > 0$) is periodically brought through the hole into the shell without touching it, then taken out of it, and so on. Determine the charge that passes through the conducting wire from the shell to ground.

**P6.11.** Three coaxial conducting hollow cylinders have radii $a = 0.5$ cm, $b = 1$ cm, and $c = 2$ cm, and equal lengths $d = 10$ m. The middle cylinder is charged with a charge $Q = 1.5 \cdot 10^{-10}$ C, and the other two are uncharged. Determine the voltages between the middle cylinder and the other two. Neglect effects at the ends of the cylinders.

**P6.12.** A charged conducting sphere of radius $b = 1$ cm and with a charge $Q = 2 \cdot 10^{-12}$ C is located at the center of an uncharged conducting spherical shell of outer radius $a = 10$ cm. The inner sphere is moved to touch the shell, and returned to its initial position. Calculate the potential of the spheres in the initial and end states for the following values of the wall thickness of the large sphere: $d = 0$ (i.e., vanishingly small), $d = 1$ cm, and $d = 5$ cm.

**P6.13.** A line charge $Q'$ is at a height $h$ above a large flat conducting surface. Determine the electric field strength along the conducting surface in the direction normal to the line charge.

**P6.14.** A point charge $Q$ is at a point $(a, b, 0)$ of a rectangular coordinate system. The half-planes ($x \geq 0, y = 0$) and ($x = 0, y \geq 0$) are conducting. Determine the electric field at a point $(x, y, 0)$, where $x > 0$ and $y > 0$.

**P6.15.** Repeat problem P6.14 for a line charge parallel to the $z$ axis.

**P6.16.** A thunderstorm cloud can be represented as an electric dipole with $\pm 10$ C of charge. The bottom part of the cloud is at $h_1 = 5$ km above the ground, and the top is $h_2 = 8$ km above the ground (Fig. P6.16). The soil is wet and can be assumed to be a good conductor. (1) Find the potential and the electric field at the surface of the earth right under the cloud. (2) Find the surface charge density at points $A$ and $B$ on the surface (Fig. P6.16), for $x = 5$ km.
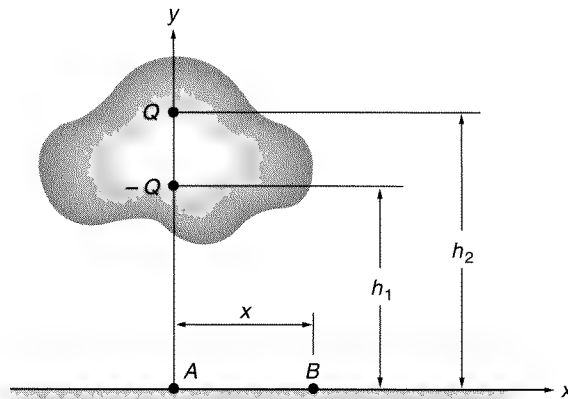


Figure P6.16 A thunderstorm cloud

**P6.17.** Find the induced charge distribution $\sigma(r)$ on the ground when a point charge $-Q$ is placed at a height $h$ above ground, assuming the ground is an infinite flat conductor. Plot your results.

**P6.18.** Repeat problem P6.17 for the case of a dipole such as the one shown in Fig. P6.16.

# 7

# Dielectrics in the Electrostatic Field

═══════════════════════════════════════════

## 7.1   Introduction

We now know that conductors change the electrostatic field by a mechanism called electrostatic induction, because any conductor has a large number of free charges that move in response to even the slightest electric field.

A wide class of substances known as *dielectrics* or *insulators* do not have free charges inside them. We might expect that, consequently, they can have no effect on the electrostatic field. This is not correct, although the mechanism by which dielectrics affect the electric field is different than in the case of conductors.

Dielectrics or insulators have many applications in electric engineering. Just as there is no electrical device without conductors, there is also no device without insulators. Therefore the analysis of dielectrics in an electrostatic field is as important as that of conductors.

## 7.2   Polarization of Dielectrics in the Electrostatic Field

Molecules of most substances behave as if electrically neutral when they are not in an electric field. We can imagine a molecule as a positive central point charge $Q$ surrounded by a spherical cloud of negative charges of total charge $-Q$ (Fig. 7.1a). This
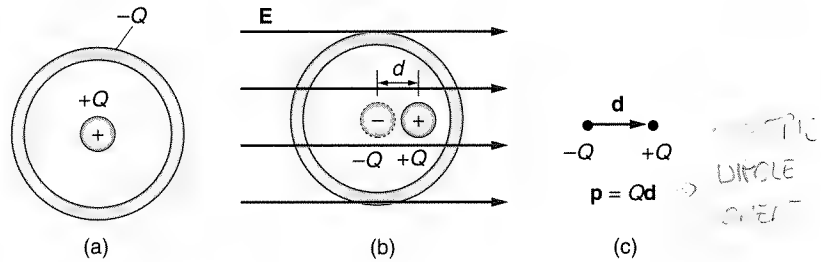
Figure 7.1 (a) Model of a nonpolar molecule, (b) the molecule in an external electric field, and (c) the electric dipole that produces the same field as the molecule in (b)

is an acceptable model, for in reality, at distances larger than a few molecular diameters, the fields of the positive and negative charges cancel out and there is no net electric field. In this rough model of a molecule, some nonelectric forces that keep the molecule spherical and symmetrical must also be present.

Assume now that we move the molecule in Fig. 7.1a into an electric field with electric field strength **E**. The field acts by a force $Q\mathbf{E}$ on the central positive charge, and by the same force, in the opposite direction, on the negatively charged cloud. Due to the forces keeping the molecule together, this will only slightly displace the central positive charge with respect to the center of the negatively charged cloud, as in Fig. 7.1b. The cloud produces the same field at points far away as if the total charge were at its center. Therefore, if we are interested in the electric field produced by the deformed molecule, we can consider it as two point charges, $Q$ and $-Q$, displaced by a small distance $d$, as in Fig. 7.1c. Two such point charges are known as an *electric dipole*.

In some substances, such as water, the molecules are electric dipoles even with no applied electric field. Such molecules are known as *polar molecules*. Those that are not dipoles in the absence of the field are termed *nonpolar molecules*. In the absence of the electric field, polar molecules are oriented at random and no electric field due to them can be observed. If a polar molecule is brought into an electric field, there are forces on the two dipole charges that tend to align the dipole with the field lines (Fig. 7.2). This alignment is more pronounced for stronger fields.

Thus for dielectrics consisting of any of the two types of molecules, the external electric field makes the substances behave as huge arrays of oriented electric dipoles. We say in such a case that the dielectric is *polarized*. The process of making a dielectric polarized is known as *polarization*.

## 7.3    The Polarization Vector

According to our model, a polarized dielectric is a vast collection of electric dipoles situated in a vacuum. If we knew the charges $Q$ and $-Q$ of the dipoles and their positions, we could evaluate the electric field strength and the scalar potential at any point. This, however, would be practically impossible due to the extremely large
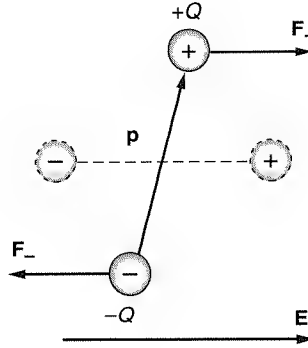
Figure 7.2 Model of a polar molecule in an external electric field

number of dipoles. For this reason we define a kind of average dipole density, a vector quantity known as the *polarization vector*.

We first need to characterize a single dipole by a vector quantity. Let $\mathbf{d}$ be the position vector of the charge $Q$ of the dipole with respect to the charge $-Q$. We define the *electric dipole moment* of the dipole (Fig. 7.3) as

$$\mathbf{p} = Q\mathbf{d} \quad (\text{C} \cdot \text{m}). \tag{7.1}$$

*(Definition of dipole moment)*

The unit of $\mathbf{p}$ is $\text{C} \cdot \text{m}$.

Consider now a small volume $dv$ of a polarized dielectric. Let $N$ be the number of dipoles per unit volume inside $dv$, and $\mathbf{p}$ be the moment of the dipoles. The polarization vector, $\mathbf{P}$, at a point inside $dv$ is defined as

$$\mathbf{P} = \frac{\sum_{dv} \mathbf{p}}{dv} = N\mathbf{p} \quad (\text{C/m}^2). \tag{7.2}$$

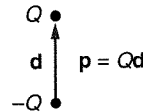*(Definition of the polarization vector)*



Figure 7.3 The dipole moment of an electric dipole is defined as the product $Q\mathbf{d}$. Note that the vector distance $\mathbf{d}$ between the two charges is adopted to be directed from the negative to the positive dipole charge.

Because the unit for the dipole moment, $\mathbf{p}$, is $C \cdot m$, the unit of $\mathbf{P}$ is $C/m^2$. Note that this is the same unit as that of the surface charge density $\sigma$.

From this definition it follows that if we know the polarization vector at a point, we can replace a small volume $dv$ (which contains a large number of dipoles) enclosing that point by a single dipole of moment

$$d\mathbf{p} = \mathbf{P}\,dv \qquad (C \cdot m). \tag{7.3}$$

*(Dipole moment of a small domain $dv$ with polarization $\mathbf{P}$)*

This expression allows us to express the scalar potential and electric field strength of a polarized dielectric as an *integral*.

Equation (7.3) can be used for the evaluation of $V$ and $\mathbf{E}$ of a polarized dielectric, but for that we need to know the expressions for $V$ and $\mathbf{E}$ of a single dipole. Consider the dipole shown in Fig. 7.4. The scalar potential at a point $P$ in the field of a dipole is obtained as the sum of potentials of the two dipole point charges:

$$V_P = \frac{Q}{4\pi\epsilon_0 r_+} + \frac{-Q}{4\pi\epsilon_0 r_-} = \frac{Q}{4\pi\epsilon_0}\left(\frac{1}{r_+} - \frac{1}{r_-}\right). \tag{7.4}$$

Because the distance $d$ between the dipole charges is always much smaller than the distance $r$ of the point $P$ from the dipole, the line segments $r, r_+,$ and $r_-$ are practically parallel. Therefore (Fig. 7.4)

$$\frac{1}{r_+} - \frac{1}{r_-} = \frac{r_- - r_+}{r_+ r_-} \simeq \frac{d\cos\theta}{r^2}, \tag{7.5}$$

so that the scalar potential at point $P$ has the form

$$V_P = \frac{Qd\,\cos\theta}{4\pi\epsilon_0 r^2} = \frac{\mathbf{p}\cdot\mathbf{u}_r}{4\pi\epsilon_0 r^2} \qquad (V), \tag{7.6}$$

where $\mathbf{u}_r$ is the unit vector directed from the dipole toward point $P$ (see Fig. 7.4). The potential of a point in the field of the dipole does not depend on $Q$ and $\mathbf{d}$ separately,
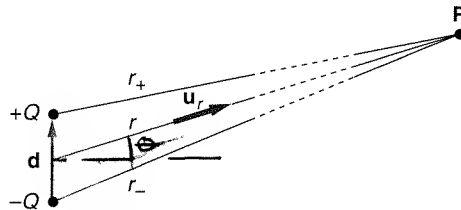


Figure 7.4 A point $P$ in the field of an electric dipole. The distance $r$ between $P$ and the dipole is much larger than the dipole size $d$

but on their product, **p**, the dipole moment. The electric field strength therefore also depends only on **p**, and not on $Q$ and **d** separately. (It is a simple matter to obtain **E** from the relation $\mathbf{E} = -\nabla V$, which is left as an exercise for the reader.)

The electric scalar potential of a polarized dielectric of volume $v$ is now obtained from Eqs. (7.3) and (7.6) as

$$V = \frac{1}{4\pi\epsilon_0} \int_v \frac{\mathbf{P} \cdot \mathbf{u}_r}{r^2} dv \qquad \text{(V)}. \tag{7.7}$$

*(Potential of a polarized dielectric body)*

When polarized, a dielectric is a source of an electric field. Consequently, the polarization of a dielectric body depends on the primary field, but also on its own polarization. It can be determined only if we know the dependence of the polarization vector on the *total* electric field strength, **E**. Experiments show that for most substances

$$\mathbf{P} = \chi_e\epsilon_0 \, \mathbf{E} \qquad \text{(}\mathbf{P}\text{ is in C/m}^2, \; \chi_e \text{ is dimensionless)}, \tag{7.8}$$

i.e., **P** at every point is proportional to **E** at that point. The constant $\chi_e$ is referred to as the *electric susceptibility* of the dielectric. If it is the same at all points, the dielectric is said to be *homogeneous*, and if it varies from point to point, the dielectric is *inhomogeneous*. Dielectrics for which Eq. (7.8) applies are known as *linear dielectrics*, and they are *nonlinear* if such a relation does not hold. For all dielectrics, $\chi_e > 0$. Only for a vacuum, $\chi_e = 0$.

*Questions and problems:*    Q7.1 to Q7.13, P7.1 to P7.3

## 7.4    Equivalent Charge Distribution of Polarized Dielectrics

A polarized dielectric can always be replaced by an equivalent volume and surface charge distribution in a vacuum. This is a very useful equivalence because we know how to determine the potential and field strength of such a charge distribution. This equivalent charge distribution can be derived from the polarization vector, **P**.

Qualitatively, when a dielectric body is brought into an electric field, as we said earlier, all the molecules become dipoles oriented in the direction of the electric field. Inside a homogeneous dielectric the fields of all the dipoles cancel out on average, because the negative part of one dipole comes close to the positive part of its identical neighbor. However, at the surface of the dielectric there will be ends of dipoles that are uncompensated. This is the extra charge that appears at the surface of a dielectric when brought into an electric field. In the case of homogeneous dielectrics, this is the *only* uncompensated charge due to polarization. Inside an inhomogeneous dielectric, there will be some net volume charge as well, because all the individual dipoles are not identical and their field does not cancel out on average anymore. Both surface and volume polarization charges can now be considered to be in a vacuum, as the rest of the dielectric does not produce any field.

The relationship between the polarization charge inside a closed surface and the polarization vector on the surface can be derived by counting the charge that passes through a surface during the polarization process (the derivation is not given in this text). The resulting expression for the polarization charge in terms of **P** is

$$Q_{\text{p in } S} = -\oint_S \mathbf{P} \cdot d\mathbf{S} \quad \text{(C).} \tag{7.9}$$

*(Polarization (excess) charge in a closed surface enclosing a polarized dielectric)*

**Example 7.1—Proof that the volume polarization charge density is zero inside a homogeneous polarized dielectric.** Consider a polarized *homogeneous* dielectric of electric susceptibility $\chi_e$, *with no volume distribution of free charges*, and a small closed surface $\Delta S$ in it. Because we have replaced the dielectric with equivalent charges in a vacuum, Gauss' law applies and the *total* charge, free and polarization, enters on the right-hand side of the formula for Gauss' law. By assumption, there are no free charges in $\Delta S$, and therefore

$$\epsilon_0 \oint_{\Delta S} \mathbf{E} \cdot d\mathbf{S} = Q_{\text{p in } \Delta S}. \tag{7.10}$$

According to Eq. (7.9), $Q_{\text{p in } \Delta S}$ can also be expressed as

$$Q_{\text{p in } \Delta S} = -\oint_{\Delta S} \mathbf{P} \cdot d\mathbf{S} = -\chi_e \epsilon_0 \oint_{\Delta S} \mathbf{E} \cdot d\mathbf{S}. \tag{7.11}$$

Since $\chi_e > 0$, Eqs. (7.10) and (7.11) can both be satisfied only if the flux of **E** through $\Delta S$ is zero. The flux of **P** through $\Delta S$ is therefore also zero. This means that *inside a homogeneous dielectric there can be no volume distribution of polarization charges, i.e., polarization charges reside only in a thin layer on the dielectric surface.*

*Questions and problems:* Q7.14 and Q7.15

## 7.5    Density of Volume and Surface Polarization Charge

Consider now an *inhomogeneous* polarized dielectric. We will show that inside such a dielectric there *is* a volume distribution of polarization charges. To determine the density of these charges, $\rho_p$, we start from Eq. (7.9). Imagine a small closed surface $\Delta S$ enclosing the point at which we wish to determine $\rho_p$. The left-hand side of Eq. (7.9) can be written as a product of $\rho_p$ and the volume $\Delta v$ enclosed by $\Delta S$. Consequently,

$$\rho_p = -\left( \frac{\oint_{\Delta S} \mathbf{P} \cdot d\mathbf{S}}{\Delta v} \right)_{\Delta v \to 0} \quad \text{(C/m}^3\text{).} \tag{7.12}$$

The expression in parentheses is known as the *divergence* of vector **P**. (For additional explanations of the concept of divergence, read Section A1.4.2 of Appendix 1 before proceeding.) It can always be evaluated from this definition in any coordinate system. In a rectangular coordinate system, the divergence of a vector **P** has the form

$$\text{div}\mathbf{P} = \frac{\partial P_x}{\partial x} + \frac{\partial P_y}{\partial y} + \frac{\partial P_z}{\partial z} \qquad (C/m^3), \tag{7.13}$$

*(Divergence in a rectangular coordinate system)*

where $P_x$, $P_y$, and $P_z$ are scalar rectangular components of the vector $\mathbf{P}$. Using the del operator, Eq. (4.20), the expression for the divergence on the right side of this equation can formally be written in a short form,

$$\text{div}\mathbf{P} = \nabla \cdot \mathbf{P}. \tag{7.14}$$

Thus the volume density of polarization charges can be written as

$$\rho_p = -\text{div}\mathbf{P} = -\nabla \cdot \mathbf{P} \qquad (C/m^3). \tag{7.15}$$

*(Volume density of polarization charges)*

Note that Eq. (7.15) is but a shorthand of Eq. (7.12), and that in a rectangular coordinate system, which we will use frequently, $\nabla \cdot \mathbf{P}$ is given by Eq. (7.13).

To determine the density of surface polarization charges, consider Fig. 7.5, showing the interface between two polarized dielectrics, 1 and 2. Apply Eq. (7.9) to the closed surface that looks like a coin, shown in the figure. There is no flux of vector $\mathbf{P}$ through the curved surface because its height approaches zero. Therefore the flux through the closed surface $\Delta S$ is given by

$$\oint_{\Delta S} \mathbf{P} \cdot d\mathbf{S} = \mathbf{P}_1 \cdot \Delta\mathbf{S}_1 + \mathbf{P}_2 \cdot \Delta\mathbf{S}_2 \qquad (C).$$

Let us adopt the reference unit vector, $\mathbf{n}$, normal to the interface, to be directed into dielectric 1 (Fig. 7.5). Then we can write $\Delta\mathbf{S}_1 = \Delta S_1\,\mathbf{n}$ and $\Delta\mathbf{S}_2 = -\Delta S_1\,\mathbf{n}$. The
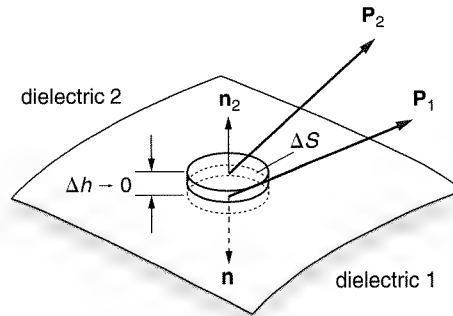
Figure 7.5 Interface between two polarized dielectrics

surface charge density is obtained if we divide the charge enclosed by $\Delta S$ by the area $\Delta S_1$ cut out of the interface by $\Delta S$. So we have

$$\sigma_p = \mathbf{n} \cdot (\mathbf{P}_2 - \mathbf{P}_1) \qquad (C/m^2). \tag{7.16}$$

*(Surface density of polarization charges on the interface between two dielectrics)*

If we know the polarization vector at all points of a dielectric, from Eq. (7.15) we can find the density of volume polarization charges (if they exist), and from the last equation we can find the density of surface polarization charges (which *always* exist). Because there are no excess charges in the rest of the dielectric, it can be disregarded. The problem of dielectric bodies in electrostatic fields is therefore reduced to that of a *distribution of charges in a vacuum,* a problem we know how to solve. What remains to be done is the determination of the polarization vector at all points. In most instances this is hard to do, but in many important cases it can be done using numerical methods.

*Questions and problems:*    Q7.16 to Q7.19, P7.4 to P7.9

## 7.6    Generalized Form of Gauss' Law: The Electric Displacement Vector

With the knowledge from the preceding section, Gauss' law can be extended to electrostatic fields with dielectric bodies.

We know that from the electrostatic-field point of view, a polarized dielectric body can be considered as a distribution of volume and surface polarization charges in a vacuum. Gauss' law is valid for a vacuum. Therefore it is straightforward to extend Gauss' law to the case of fields with dielectrics: simply add the polarization charge to the free charge enclosed by $S$. Consequently, Gauss' law in Eq. (5.4) becomes

$$\oint_S \mathbf{E} \cdot d\mathbf{S} = \frac{Q_{\text{free in } S} + Q_{\text{polarization in } S}}{\epsilon_0}. \tag{7.17}$$

Usually, this generalized Gauss' law is written in a different form. First, the polarization charge in $S$ is represented as in Eq. (7.9). Note that the surface $S$ is the same for the integral on the left-hand side of Eqs. (7.17) and (7.9). We can, therefore, multiply Eq. (7.17) by $\epsilon_0$, move the integral representing $Q_{\text{polarization in } S}$ to the left-hand side of Eq. (7.17), and use just one integral sign. The result of this manipulation is

$$\oint_S (\epsilon_0 \mathbf{E} + \mathbf{P}) \cdot d\mathbf{S} = Q_{\text{free in } S} \qquad (C). \tag{7.18}$$

This is a very interesting result: the flux of the sum of the vectors $\epsilon_0 \mathbf{E}$ and $\mathbf{P}$ through any closed surface $S$ is equal to the total *free* charge enclosed by $S$. The form of Gauss' law (7.18) is more convenient than that of (7.17) because the only charges we can influence directly are free charges.

To simplify Eq. (7.18), we define the *electric displacement vector*, **D**, as

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \quad (\text{C/m}^2).$$ (7.19)

*(Definition of the electric displacement vector)*

With this definition, the generalized Gauss' law takes the final form:

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = Q_{\text{free in S}} \quad (\text{C}).$$ (7.20)

*(Generalized Gauss' law)*

The expression in Eq. (7.19) is the most general definition of the electric displacement vector. If the dielectric is linear (as most, but not all, dielectrics are), vector **D** can be expressed in terms of the electric field strength, **E**:

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \epsilon_0 \chi_e \mathbf{E} = \epsilon_0 (1 + \chi_e) \mathbf{E} = \epsilon_0 \epsilon_r \mathbf{E} = \epsilon \mathbf{E} \quad (\text{C/m}^2),$$ (7.21)

*(Electric displacement vector in linear dielectrics)*

where

$$\epsilon_r = (1 + \chi_e) \quad \text{(dimensionless)}$$ (7.22)

*(Definition of relative permittivity—linear dielectrics only)*

is known as the *relative permittivity* of the dielectric, and

$$\epsilon = \epsilon_r \epsilon_0 \quad (\text{F/m}).$$ (7.23)

*(Definition of permittivity—linear dielectrics only)*

as the *permittivity* of the dielectric.

Because the electric susceptibility, $\chi_e$, is always greater than zero, the relative permittivity, $\epsilon_r$, is always greater than unity. The most frequent values of $\epsilon_r$ are between 2 and about 10, but there are dielectrics with much higher relative permittivities. For example, distilled water (which is a dielectric) has relative permittivity of about 80 (this is because its molecules are polar molecules). A table of values of relative permittivities for some common dielectrics is given in Appendix 4.

**Example 7.2—Electric field in a *pn* diode.** A *pn* diode, sketched in Fig. 7.6, is a fundamental semiconductor device and is a part of all bipolar transistors. Unlike in a metal, where electrons are the only charge carriers, in a semiconductor diode both negative and positive free charges are responsible for current flow when the diode is biased. The semiconductor material
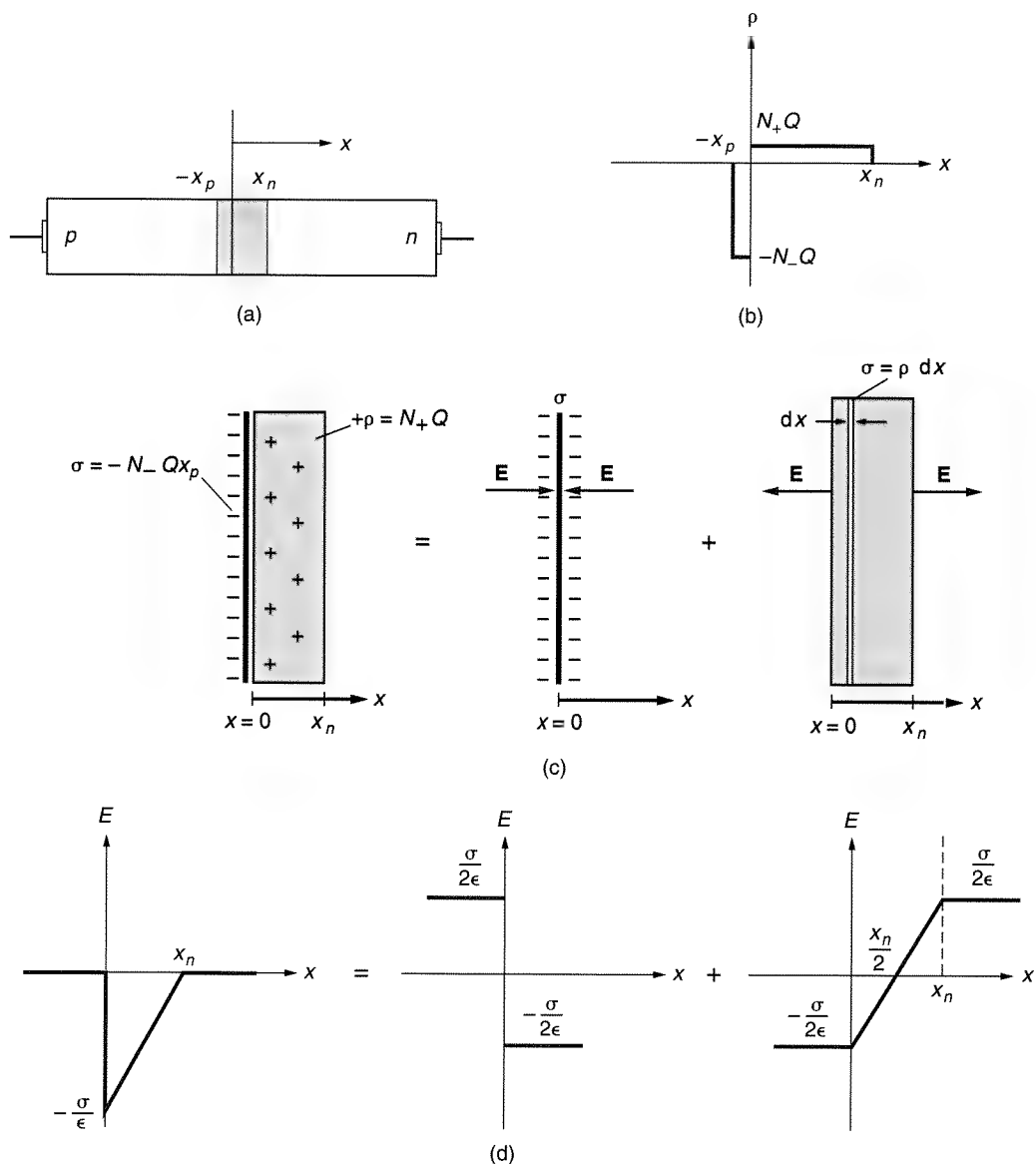
Figure 7.6 (a) Sketch of a *pn* diode and (b) its approximate charge density profile. (c) A diode can be approximated by a sheet of negative surface charge and a bulk of positive volume charge. (d) Superposition of the individual fields of the two charge distributions from (c) gives the final field distribution in the diode.

has a permittivity $\epsilon$ (for silicon $\epsilon_r = 11$, and for gallium arsenide $\epsilon_r = 13$), and if it is pure it behaves as a dielectric. When certain impurities called *dopants* are added to the material, it becomes conductive. The $p$ region of the diode is a doped semiconductor material that has $p$ *positive* free charge carriers per unit volume. This part is in physical contact with the $n$ region, which has $n$ *negative* free charge carriers per unit volume.

When the two parts are put together but not biased, the negative charge carriers (electrons) diffuse into the neighboring $p$ region. Positive charge carriers ("holes" with a charge equal to that of an electron) diffuse into the neighboring $n$ region. (The diffusion process is similar to the diffusion of two different gases through a thin membrane, except that the diffused charge carriers remain in the immediate vicinity of the boundary surface.) Because the negative charge carriers move into the region from which positive charge carriers partly left, leaving behind negatively charged atoms, there will be a surplus of negative charge in this thin layer of the $p$ region. Similarly, there will be a surplus of positive charges in the adjoining thin layer of the $n$ region.

These two charged layers produce an electric field (as in a parallel-plate capacitor), resulting in an electric force on free charge carriers that opposes the diffusion process. This electric force eventually (actually, in a very short time) stops the diffusion of free charge carriers. Thin layers on both sides of the boundary surface are thus depleted of their own free charge carriers. These two layers are known as the *depletion region*. Consequently, the depletion region finds itself between the $p$ and $n$ undepleted regions, and contains two layers of equal and opposite charges. Let the number of *positive* charges per unit volume in the $n$ region be $N_+$, and the number of *negative* charges in the $p$ region be $N_-$. The volume densities of charge in the two layers of the depletion region are $\rho_+ = N_+ Q$ (in the $n$ part), and $\rho_- = -N_- Q$ (in the $p$ part), where $Q$ is the absolute value of the electron charge.

If the diode is not biased (its two terminals are left open), the opposite charges on the two sides of the junction are of equal magnitude. Therefore the thicknesses of the two charged layers, $x_p$ and $x_n$, are connected by the relation $N_- x_p = N_+ x_n$. Usually the diode is made so that the $n$ side of the junction has a much larger concentration of diffused negative free charge carriers than the other, that is, $N_- \gg N_+$. This means that $x_n \gg x_p$. Such a junction is called a one-sided step junction, and its charge concentration profile is sketched in Fig. 7.6b. This tells us that the width of the depletion layer on the $p$ side can be neglected to the first order, i.e., this charged layer can be approximated by a negatively charged sheet of a surface charge density $\sigma = N_- Q / x_p$, Fig. 7.6c. On the $n$ side, the depletion layer is effectively a uniform volume charge density (that is, $N_+$ is coordinate-independent). We already know from Example 5.3 what the field of the negative surface-charge sheet is, and it is shown in the middle of Fig. 7.6c.

What is the electric field of a volume charge, such as the one on the right in Fig. 7.6c? Outside the charged layer, it is equal to the field of a charged sheet of the same *total* charge:

$$E_{\text{outside}} = \frac{\sigma}{\epsilon} = \frac{\rho x_n}{2\epsilon} = \frac{N_+ Q x_n}{2\epsilon}. \tag{7.24}$$

Inside the volume charge, we can apply Gauss' law to a thin slice $dx$ wide, as indicated on the right in Fig. 7.6c, which contains $\rho\, dx$ surface charge. It is left to the reader to show that integration of the field resulting from all the slices between $0$ and $x_n$ gives the following expression for the electric field inside the volume charge density as a function of the $x$ coordinate:

$$E_{\text{inside}} = \frac{\rho}{\epsilon}\left(x - \frac{x_n}{2}\right) = \frac{N_+ Q x_n}{\epsilon}\left(x - \frac{x_n}{2}\right). \tag{7.25}$$

DIRECTED IN THE +X DIRECTION IF X>Xₙ, AND IN THE −X DIRECTION IF X < Xₙ

This expression is shown graphically on the ~~left~~ **RIGHT** in Fig. 7.6d. Using the principle of superposition, we can now add the field of the negative surface charge (in the middle of Fig. 7.6d) to the field of the positive volume charge we found (on the ~~left~~ **RIGHT** in Fig. 7.6d) to get the field profile of a *pn* diode, shown on the ~~right~~ **LEFT** in Fig. 7.6d. It is left to the reader as an exercise to sketch the potential distribution inside a diode.

*Questions and problems:*   Q7.20 and Q7.21, P7.10

## 7.7   Electrostatic Boundary Conditions

In inhomogeneous media consisting of several homogeneous parts there is, obviously, an abrupt change in some quantities describing the field on the two sides of boundaries. For example, if on such a boundary there is a surface polarization charge, it is a source of the electric field component directed in opposite directions on the two sides of the boundary; consequently, the total electric field must have a different direction and magnitude on the two sides of the boundary.

Such abrupt changes of any quantity describing the field must satisfy basic field equations and definitions. Specialized field equations describing this behavior, more precisely connecting the values of any field quantity on two sides of a boundary surface, are known as *boundary conditions*. What are boundary conditions needed for? Note that they represent, in fact, fundamental equations of the electrostatic field specialized to boundary surfaces. Therefore in a medium consisting of several dielectric bodies, the field transition from one body to the adjacent body through a boundary surface *must* be as dictated by the boundary conditions. Otherwise this could not be a real electric field because it would not satisfy the field equations *everywhere*. Note that this is true for all boundary conditions we introduce in later chapters.

Let us apply first the law of conservation of energy of the electrostatic field, Eq. (4.7), to the narrow rectangular contour $\Delta C$ in Fig. 7.7. Because the length of the shorter sides approaches zero, the contribution to the line integral of **E** along them is
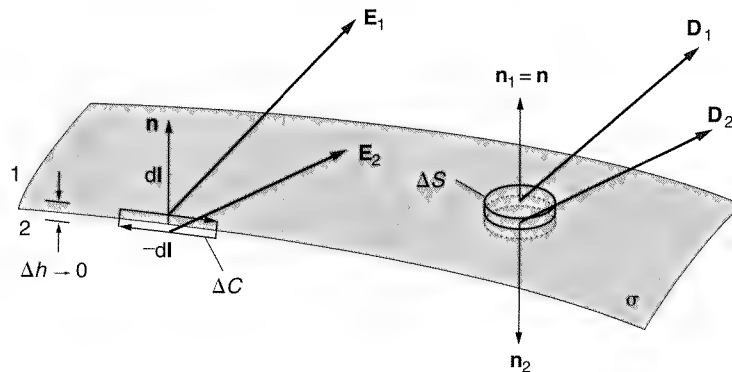


Figure 7.7 Boundary between two media. A narrow rectangular contour is used in the law of conservation of energy and a coinlike closed surface is used in Gauss' law for deriving boundary conditions for vectors **E** and **D**, respectively.

zero. Along the two longer sides, the contribution is $(\mathbf{E}_1 \cdot d\mathbf{l}_1 + \mathbf{E}_2 \cdot d\mathbf{l}_2)$. The scalar products are simply tangential components of the two electric field strength vectors, which we denote by the subscript "t." Because $d\mathbf{l}_2 = -d\mathbf{l}_1$, the boundary condition for the tangential components of vector $\mathbf{E}$ is

$$E_{1t} = E_{2t} \quad \text{(valid in general).} \tag{7.26}$$

*(Boundary condition for tangential components of vector $\mathbf{E}$)*

Note that no other assumptions are needed to derive this condition except Eq. (4.7). Consequently, it is valid for all cases of the electrostatic field. We will see that it is valid also for the general case of a time-varying electromagnetic field.

Now let us apply Gauss' law, Eq. (7.20), to the small cylindrical coinlike surface in Fig. 7.7. Let there be a surface charge $\sigma$ on the boundary inside the surface. There is no flux of vector $\mathbf{D}$ through the curved surface because its height is vanishingly small. The flux through the two cylinder bases is $D_{1n} \Delta S$ (the outward flux) and $-D_{2n} \Delta S$ (the inward flux), both with respect to the reference unit vector $\mathbf{n}$ directed into dielectric 1, where the subscript "n" denotes the normal component. The enclosed charge being $\sigma \Delta S$, the generalized Gauss' law yields

$$\mathbf{D}_1 \cdot \mathbf{n} - \mathbf{D}_2 \cdot \mathbf{n} = \sigma, \text{ or } D_{1n} - D_{2n} = \sigma \quad \text{(valid in general).} \tag{7.27}$$

*(Boundary condition for normal component of vector $\mathbf{D}$; unit vector normal, $\mathbf{n}$, directed into medium 1)*

In the special case when there is no surface charge on the boundary, this becomes

$$D_{1n} = D_{2n} \quad \text{(no free surface charges on boundary).} \tag{7.28}$$

Another important case is the boundary between a conductor and a dielectric. Let the dielectric be medium 1, and the conductor be medium 2. We know that there is no field inside a conductor. Therefore $D_{2n} = 0$, and Eq. (7.27) becomes

$$D_n = \sigma \quad \text{(on boundary of dielectric and conductor).} \tag{7.29}$$

Note that this is essentially the same equation as Eq. (6.5). We will see that Eqs. (7.27) to (7.29) are also valid in general, and not only for electrostatic fields.

*Questions and problems:*   Q7.22 to Q7.24, P7.11 to P7.15

## 7.8    Differential Form of Generalized Gauss' Law

The generalized Gauss' law in Eq. (7.20) can be transformed into a differential equation, known as the differential form of Gauss' law. To obtain this differential equation, let us apply Eq. (7.20) to a small volume $\Delta v$ enclosed by a surface $\Delta S$, and divide both sides of the equation by $\Delta v$. The right side then becomes simply the volume charge density, $\rho$, inside $\Delta S$. The left side becomes the same as the expression in Eq. (7.12), with **P** substituted by **D**. We know that this expression is the divergence of vector **D**. So we obtain

$$\mathrm{div}\mathbf{D} = \rho. \tag{7.30}$$

*(Differential form of generalized Gauss' law)*

Since the divergence of **D** is a combination of derivatives of the components of **D**, this is indeed a differential equation in three unknowns, the three scalar components of vector **D**. It is known as a partial differential equation because partial derivatives, with respect to individual coordinates, enter into the equation. We will see that the basic equations of the electromagnetic field, Maxwell's equations, are a set of four partial differential equations. Equation (7.30) is one of these four equations.

## 7.9    Poisson's and Laplace's Equations: The Laplacian

The potential at a point is related to the volume charge density at that point by a differential equation known as *Poisson's equation*. A special case of Poisson's equation for the case when the volume charge density is zero is called *Laplace's equation*. The derivation of these equations is quite simple.

We know that we can always represent vector **E** as $\mathbf{E} = -\mathrm{grad}\, V = -\nabla V$. For linear media, therefore, $\mathbf{D} = -\epsilon\, \mathrm{grad}\, V = -\epsilon\, \nabla V$, so that from the generalized form of Gauss' law, Eq. (7.13), we obtain

$$\mathrm{div}(\epsilon\, \mathrm{grad} V) = \nabla \cdot (\epsilon \nabla V) = -\rho. \tag{7.31}$$

This is the most general form of Poisson's equation. For the frequent case of a homogeneous dielectric ($\epsilon$ the same at all points), Eq. (7.31) becomes

$$\mathrm{div}(\mathrm{grad} V) = \nabla \cdot (\nabla V) = -\frac{\rho}{\epsilon}. \tag{7.32}$$

*(Poisson's equation)*

Laplace's equation is obtained from Eqs. (7.31) and (7.32) if we set $\rho = 0$:

$$\mathrm{div}(\epsilon\, \mathrm{grad} V) = \nabla \cdot (\epsilon \nabla V) = 0 \tag{7.33}$$

for a general, inhomogeneous dielectric with no free charges, and

$$\mathrm{div}(\mathrm{grad}\,V) = \nabla \cdot (\nabla V) = 0 \qquad (7.34)$$

*(Laplace's equation)*

for a homogeneous dielectric with no free charges.

The operator $\mathrm{div}(\mathrm{grad}) = \nabla \cdot \nabla$ is known as *Laplace's operator,* or the *Laplacian,* and is denoted briefly as $\Delta$ or $\nabla^2$. It is a simple matter to show that, in a rectangular coordinate system, Laplace's operator has the form

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \qquad (7.35)$$

*(Laplacian operator in rectangular coordinate system)*

As an important example, if the volume charge distribution in a region is a function of a single rectangular coordinate, for example of $x$, $V$ is then also a function of $x$ only. Poisson's equation becomes

$$\frac{d^2 V(x)}{dx^2} = -\frac{\rho(x)}{\epsilon}. \qquad (7.36)$$

This equation is used often, for example, in the analysis of semiconductor devices including diodes, transistors, and capacitors.

**Example 7.3—The *pn* Diode Revisited.** In this example, we use Poisson's equation to find the potential distribution in a *pn* diode, using the one-sided step junction approximation from Example 7.2. Poisson's equation for the $p$ side of the junction can be written as

$$\frac{d^2 V}{dx^2} = -\left(-\frac{QN_-}{\epsilon_0 \epsilon_r}\right) = \frac{QN_-}{\epsilon_0 \epsilon_r}, \qquad (7.37)$$

and for the $n$ side as

$$\frac{d^2 V}{dx^2} = -\frac{QN_+}{\epsilon_0 \epsilon_r}. \qquad (7.38)$$

However, in the one-sided step approximation, the width of the depletion layer on the $p$ side is negligible, so we only need to solve Eq. (7.38). We first integrate once with respect to $x$ from 0 to $x$. We need one boundary condition to determine the integration constant in this step. We know that there is no electric field outside of the depletion region, so the boundary condition is $dV/dx = 0$ at $x = x_n$. Integrating Eq. (7.38) once therefore yields

$$\frac{dV}{dx} = -\frac{QN_+}{\epsilon_0 \epsilon_r}(x - x_n). \qquad (7.39)$$

Because we know that $\mathbf{E} = -(dV/dx)\mathbf{u}_x$, we can rearrange terms in Eq. (7.39) to obtain the same expression for the electric field as the one shown graphically in Fig. 7.6d. To get the potential, we integrate another time. Let us adopt as the boundary condition that the potential is zero at $x = x_n$ (we know that we can adopt it to be zero at any point). We thus obtain

$$V(x) = -\frac{QN_+x_n^2}{2\epsilon_0\epsilon_r}\left(1 - \frac{x}{x_n}\right)^2. \tag{7.40}$$

As this potential exists inside the diode even when its terminals are not connected to an external voltage source, it is called the *built-in potential*.

When a bias is applied to a diode, it changes the width of the depletion layer. If we connect the diode *p* region to the positive output of a voltage source and the *n* side to the negative one, the depletion layer gets narrower, making it easier for free charges to flow through it. This is called *forward bias*. If the diode terminals are connected the other way, the depletion layer becomes thicker and current flow is disabled. This is called *reverse bias*. If an ac voltage is applied to the diode, in one half of the cycle the diode will conduct and in the other half there will be no current. Therefore a diode is a *rectifier*.

*Questions and problems:*   P7.16 to P7.22

## 7.10   Some Practical Electrical Properties of Dielectrics

Applications of dielectrics in electrical engineering are hardly possible without knowing their electrical properties. We briefly mention here some of these properties.

In addition to relative permittivity, two more properties need particular attention. The first is the *dielectric strength* of a dielectric. This is the largest magnitude of the electric field that can exist in a dielectric without damaging it. If the field magnitude is greater than the dielectric strength of the dielectric, *dielectric breakdown* occurs (the dielectric burns, cracks, ionizes, and becomes conductive, becomes very lossy, etc.).

The typical value of the dielectric strength for air is about $3 \cdot 10^6$ V/m, or 30 kV/cm. For liquid and solid dielectrics the electric field strength ranges from about $15 \cdot 10^6$ V/m to about $40 \cdot 10^6$ V/m. Values of the dielectric strength of some common dielectrics are given in Appendix 4.

Another important property of dielectrics is loss that produces heat. Most dielectrics have a very small number of free charges, so that resistive (Joule's) losses in them due to time-constant fields (except for very large field magnitudes) are usually negligible. In time-varying fields, however, there is a new type of loss, known as the *polarization loss*, that is much larger than Joule's losses. Qualitatively, the time-varying electric field induces time-varying dipoles in the dielectric, which start to vibrate more vigorously due to these oscillations. This vibration is heat, i.e., it represents losses to the field polarizing the dielectric.

*Questions and problems:*   Q7.25 to Q7.27, P7.23

## 7.11   Chapter Summary

1. If introduced into an electrostatic field, all dielectrics can be visualized as a vast ensemble of small electric dipoles situated in a vacuum. We say that such a dielectric is polarized.

2. The polarization of a dielectric at any point is described by the polarization vector, $\mathbf{P}$, representing a vector density of dipole moments at that point. The dipole moment of a dipole of charges $Q$ and $-Q$ separated by a distance $\mathbf{d}$ (directed from $-Q$ to $Q$) is defined as $\mathbf{p} = Q\mathbf{d}$.

3. The polarized dielectric can further be considered as an equivalent distribution of volume and surface charges, known as *polarization charges*. These two charge densities are determined in terms of the polarization vector, $\mathbf{P}$. The rest of the dielectric has no effect whatsoever on the field and can be removed. The polarization charges must, therefore, be considered to be situated in a vacuum.

4. The vector quantity $\mathbf{D} = (\epsilon_0\mathbf{E} + \mathbf{P})$ has a simple and useful property: its flux through any closed surface equals the total free charge inside the surface. This equation is known as the *generalized Gauss' law*, and vector $\mathbf{D}$ as the *electric displacement vector*.

5. The generalized Gauss' law can also be written in the form of a differential equation, $\nabla \cdot \mathbf{D} = \rho$. This is known as the differential form of Gauss' law.

6. There is a simple differential relationship between the potential function at a point and volume charge density at that point, known as the Poisson equation, $\nabla \cdot \epsilon \nabla V = -\rho$. Its special form, when there are no volume charges, is known as Laplace's equation, $\nabla \cdot \epsilon \nabla V = 0$.

### QUESTIONS

**Q7.1.** At a point of a polarized dielectric there are $N$ dipoles per unit volume. Each dipole has a moment $\mathbf{p}$. What is the polarization vector at that point?

**Q7.2.** A body is made of a linear, homogeneous dielectric. Explain what this means.

**Q7.3.** What is the difference between an inhomogeneous linear dielectric and a homogeneous nonlinear dielectric?

**Q7.4.** Why is $\chi_e = 0$ for a vacuum?

**Q7.5.** Are there substances for which $\chi_e < 0$? Explain.

**Q7.6.** An atom acquires a dipole moment proportional to the electric field strength $\mathbf{E}$ of the external field, $\mathbf{p} = \alpha\mathbf{E}$ ($\alpha$ is often referred to as the *polarizability*). Determine the electric force on the atom if it is introduced into a *uniform* electric field of intensity $\mathbf{E}$.

**Q7.7.** Answer question Q7.6 for the case in which the atom is introduced into the field of a point charge $Q$. Determine only the direction of the force, not its magnitude.

**Q7.8.** A small body—either dielectric or conducting—is introduced into a nonuniform electric field. In which direction (qualitatively) does the force act on the body?

**Q7.9.** Two point charges are placed near a piece of dielectric. Explain why Coulomb's law cannot be used to determine the *total* force on the two charges.

**Q7.10.** A small charged body is placed near a large dielectric body. Will there be a force acting between the two bodies? Explain.

**Q7.11.** A closed surface $S$ situated in a vacuum encloses a total charge $Q$ *and* a polarized dielectric body. Using a sound physical argument, prove that in this case also the flux of the electric field strength vector $\mathbf{E}$ through $S$ is $Q/\epsilon_0$.

**Q7.12.** Arbitrary pieces of dielectrics and conductors carrying a total charge $Q$ are introduced through an opening in a hollow, uncharged metal shell. The opening is then closed. Using a physical argument and Gauss' law for a vacuum, prove that the charge appearing on the outer surface of the shell is exactly equal to $Q$.

**Q7.13.** A positive point charge is placed in air near the interface of air and a liquid dielectric. Will the interface be deformed? If you think it will be deformed, then will it raise or sink? What if the charge is negative?

**Q7.14.** Explain in your own words why Eqs. (7.10) and (7.11) imply that the flux of $\mathbf{E}$ through a closed surface $\Delta S$ is zero.

**Q7.15.** Electric dipoles are arranged along a line (possibly curved) so that the negative charge of one dipole coincides with the positive charge of the next. Describe the electric field of this arrangement of dipoles.

**Q7.16.** Write Eq. (7.16) for the interface of a dielectric and a vacuum. For case (1) assume the dielectric to be medium 1, and for case (2) medium 2.

**Q7.17.** Is there a pressure of electrostatic forces acting on a boundary surface between two different dielectrics situated in an electrostatic field? Explain.

**Q7.18.** Prove that the total polarization charge in any piece of a dielectric material is zero.

**Q7.19.** A point charge $Q$ is placed inside a spherical metal shell, a distance $d$ from its center. In addition, the shell is filled with an inhomogeneous dielectric. Determine the electric field strength outside the shell.

**Q7.20.** Does Eq. (7.18) mean exactly the same as Eq. (7.17)? Explain.

**Q7.21.** Can the relative permittivity of a dielectric be less than one, or negative? Explain.

**Q7.22.** Can you find an analogy between properly connecting sleeves to a jacket, and using boundary conditions in solving electrostatic field problems? Describe.

**Q7.23.** Prove that a charged conductor situated in an inhomogeneous but linear dielectric has a potential proportional to its charge. [*Hint:* consider the polarized dielectric as an aggregate of dipoles situated in a vacuum.]

**Q7.24.** Discuss question Q7.23 for a case in which the dielectric is not linear.

**Q7.25.** What is the unit of dielectric strength of a dielectric?

**Q7.26.** Explain how $30\,\mathrm{kV/cm}$ is the same as $3 \cdot 10^6\,\mathrm{V/m}$.

**Q7.27.** Are polarization losses in a dielectric the same as resistive Joule's losses? Explain.

## PROBLEMS

**P7.1.** Using the relation $\mathbf{E} = -\nabla V$, determine the spherical components $E_r$, $E_\theta$, and $E_\phi$ of the electric field strength of the electric dipole in Fig. 7.4.

**P7.2.** Determine the electric force on a dipole of moment $\mathbf{p}$ located at a distance $r$ from a point charge $Q_0$, if the angle between $\mathbf{p}$ and the direction from the charge is arbitrary.

**P7.3.** An atom acquires a dipole moment proportional to the electric field strength $\mathbf{E}$ of the external field, $\mathbf{p} = \alpha\mathbf{E}$. Determine the force on the dipole if it is introduced into the field of a point charge $Q$ at a distance $r$ from the charge.

**P7.4.** A homogeneous dielectric sphere is polarized uniformly over its volume. The polarization vector is $\mathbf{P}$. Determine the distribution of the polarization charges inside and on the surface of the sphere.

**P7.5.** A thin circular dielectric disk of radius $a$ and thickness $d$ is permanently polarized with a dipole moment per unit volume $\mathbf{P}$, parallel to the axis of the disk that is normal to its plane faces. Determine the electric field strength and the electric scalar potential along the disk axis. Plot your results.

**P7.6.** Determine the density of volume polarization charges inside a linear but inhomogeneous dielectric of permittivity $\epsilon(x, y, z)$ at a point where the electric field strength is $\mathbf{E}$. There is no volume distribution of free charges inside the dielectric.

**P7.7.** The permittivity of an infinite dielectric medium is given as the following function of the distance $r$ from the center of symmetry: $\epsilon(r) = \epsilon_0(1 + a/r)$. A small conducting sphere of radius $R$, carrying a charge $Q$, is centered at $r = 0$. Determine and plot the electric field strength and the electric scalar potential as functions of $r$. Determine the volume density of polarization charges.

**P7.8.** A conducting sphere of radius $a$ carries a charge $Q$. Exactly one half of the sphere is pressed into a dielectric half-space of permittivity $\epsilon$. Air is above the dielectric. Determine the free and polarization surface charge density on the sphere and in the dielectric.

**P7.9.** Repeat problem P7.8 for a circular cylinder of radius $a$ with charge $Q'$ per unit length.

**P7.10.** A small spherical charged body with a charge $Q = -1.9 \cdot 10^{-9}$ C is located at the center of a spherical dielectric body of radius $a$ and relative permittivity $\epsilon_r = 3$. Determine the vectors $\mathbf{E}$, $\mathbf{P}$, and $\mathbf{D}$ at all points, volume and surface density of polarization charges, and the potential at all points. Is it possible to determine the field and potential outside the dielectric body without solving for the field inside the body? Explain.

**P7.11.** What is $\mathbf{E}$ equal to in a needlelike air cavity inside a homogeneous dielectric of permittivity $\epsilon$ if the cavity is parallel to the electric field vector $\mathbf{E}_d$ inside the dielectric (Fig. P7.11)?
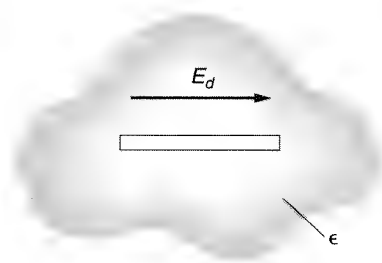


Figure P7.11 A needlelike cavity

**P7.12.** What is $\mathbf{E}$ equal to in a disklike air cavity with faces normal to the electric field vector $\mathbf{E}_d$ inside a homogeneous dielectric of permittivity $\epsilon$ (Fig. P7.12)?
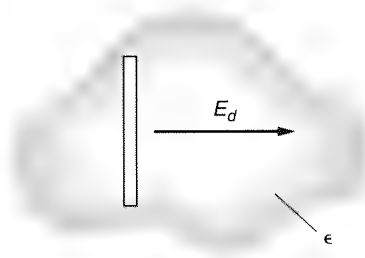
Figure P7.14 A disklike cavity

**P7.13.** At a point of the boundary surface between dielectrics of permittivities $\epsilon_1$ and $\epsilon_2$, the electric field strength vector in medium 1 makes an angle $\alpha_1$ with the normal to the boundary, and that in medium 2 an angle $\alpha_2$. Prove that $\tan\alpha_1 / \tan\alpha_2 = \epsilon_1/\epsilon_2$.

**P7.14.** A dielectric slab of permittivity $\epsilon = 2\epsilon_0$ is situated in a vacuum in an external uniform electric field $\mathbf{E}$ so that the field lines are perpendicular to the faces of the slab (Fig. P7.14). Sketch the lines of the resulting vectors $\mathbf{E}$ and $\mathbf{D}$.
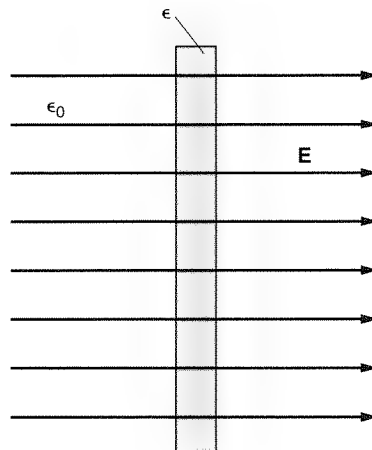


Figure P7.14 Field lines normal to dielectric slab
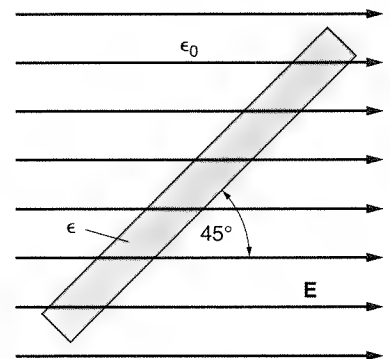
Figure P7.15 Field lines oblique to dielectric slab

**P7.15.** Repeat problem P7.14 assuming that the dielectric slab is at an angle of 45 degrees with respect to the lines of the external electric field (Fig. P7.15).

**P7.16.** One of two very large parallel metal plates is at a zero potential, and the other at a potential $V$. Starting from Laplace's equation, determine the potential, and hence the electric field strength, at all points.

**P7.17.** Two concentric spherical metal shells, of radii $a$ and $b$ ($b > a$), are at potentials $V$ (the inner shell) and zero. Starting from Laplace's equation in spherical coordinates, determine the potential, and hence the electric field strength, at all points. Plot your results.

**P7.18.** The charge density at all points between two large parallel flat metal sheets is $\rho_0$. The sheets are $d$ apart. One of the sheets is at a zero potential, and the other at a potential $V$. Find the potential at all points between the plates starting from Poisson's equation. Plot your result.

**P7.19.** Repeat problem P7.18 if the charge density between the plates is $\rho(x) = \rho_0 x/d$, $x$ being a coordinate normal to the plates, with the origin at the zero-potential plate. Plot your result and compare to problem P7.18.

**P7.20.** Repeat problem P7.19 if the origin is at the plane of symmetry of the system.

**P7.21.** Two long coaxial cylindrical thin metal tubes of radii $a$ and $b$ ($b > a$) are at potential zero (the outer tube) and $V$. Starting from Laplace's equation in cylindrical coordinates, determine the potential between the cylinders, and hence the electric field strength.

**P7.22.** Prove that if $V_1$ and $V_2$ are solutions of Laplace's equation, their product is not generally a solution of that equation.

**P7.23.** The radii of conductors of a coaxial cable with air dielectric are $a$ and $b$ ($b > a$). Determine the maximum value of the potential difference between the conductors for which a complete breakdown of the air dielectric does not occur. The dielectric strength of air is $E_0$.

# 8

# Capacitance and
# Related Concepts

## 8.1   Introduction

Capacitors consist of two metal bodies, known as the *capacitor electrodes*, charged with equal charges of opposite sign. They are characterized by a quantity known as the capacitance. Capacitors are of fundamental importance in electrical engineering and are commonly used by most engineers. Other important concepts related to the capacitance, however, are less widely understood. In this chapter, we examine electric coupling and shielding as phenomena closely related to the topic of capacitance.

## 8.2   Capacitors and Capacitance

Consider first a conductive body with charge $Q$ far away from any other charges. Assume that the potential of the body is $V$. If the charge on the body is changed to $kQ$, where $k$ is any real number, what does the potential of the body become?

The surface charge density on the body, $\sigma$, is such that the electric field is zero inside the body. The electric field at any point on the surface is given by $\mathbf{E} = (\sigma/\epsilon)\mathbf{n}$. When the body is charged to $kQ$, the electric field inside the body has to remain zero, so the surface charge density at every point *must* be $k\sigma$. This means that the new

104

electric field strength at all points will be $k\mathbf{E}$. The potential then also has to increase by a factor of $k$, so the new potential is $kV$. The conclusion is that *the charge of a conductive body and its potential are proportional to each other.* This proportionality is written as

$$Q = CV, \qquad \text{or} \qquad C = \frac{Q}{V} \qquad [C \text{ is in farads (F)}]. \tag{8.1}$$

*(Definition of capacitance of an isolated body)*

The constant $C$ does not depend on $Q$ or $V$, but only on the shape and size of the body and on the materials surrounding it. It is called the *capacitance* of an isolated body. To determine it, we need to know its potential in terms of a given charge. For example, for a metal ball of radius $a$ in a vacuum, we get

$$C = \frac{Q}{V_{\text{ball}}} = 4\pi \epsilon_0 a \qquad \text{(F)}. \tag{8.2}$$

The unit for capacitance is the *farad* (abbreviated F). It is equal to coulomb/volt.

**Example 8.1—Capacitance of the earth.** Let us use the formula for the capacitance of an isolated conducting sphere to find the capacitance of the earth:

$$C_{\text{earth}} = 4\pi \epsilon_0 R_{\text{earth}} = \frac{1}{9 \cdot 10^9} 6.37 \cdot 10^6 \simeq 0.708 \cdot 10^{-3} \, \text{F} < 1 \, \text{mF}. \tag{8.3}$$

The capacitance of the whole earth is much less than a farad. Obviously, the farad is a very large unit. The values of capacitance of commonly used capacitors are from a few pF to a few $\mu$F.

As already mentioned, a system consisting of two conductive bodies charged with equal charges of opposite sign is called a *capacitor*, shown in Fig. 8.1. The metal
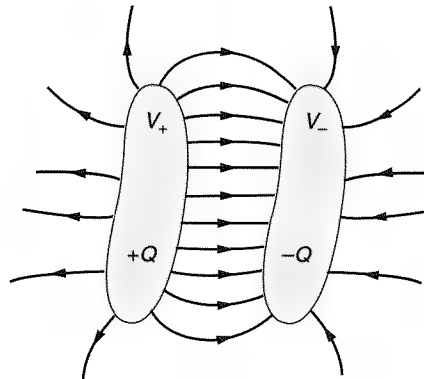


Figure 8.1 A capacitor consists of two bodies charged with equal amounts of charge, but of opposite sign.

bodies are the capacitor electrodes. Capacitors can have different shapes. Following the same reasoning as for an isolated metal body, we can conclude that the charge on the reference capacitor electrode, e.g., $Q$, is proportional to the potential difference between the two electrodes (the proof is left as an exercise for the reader; see Q8.3). This is written as

$$Q = C(V_Q - V_{-Q}), \quad \text{or} \quad C = \frac{Q}{V_Q - V_{-Q}} \quad [C \text{ is in farads (F)}]. \quad (8.4)$$
$$\text{(Definition of the capacitance of a capacitor)}$$

The constant $C$ is the capacitance of the capacitor. It depends on the shape, size, and position of the electrodes and on the properties of the dielectric between them. Usually the capacitance does not depend on the charge $Q$ on the electrodes, nor does it depend on the voltage $V_{+-} = V_Q - V_{-Q}$ between them. Such capacitors are *linear capacitors*. For *nonlinear capacitors* these conditions are not satisfied. For example, in a semiconductor device known as a varactor diode, the capacitance of the diode depends on the voltage applied to its terminals.

Although parallel and series connections of capacitors are familiar from circuit theory, we repeat them here from the field-theory point of view. We will see that some conditions implicit in the definitions of the equivalent capacitor in the two cases cannot be seen from circuit theory.

**Example 8.2—Parallel connection of capacitors.** Consider a parallel connection of capacitors as in Fig. 8.2. Although it might seem a bit strange, this is just a form of the capacitor shown in Fig. 8.1. We have two terminals connected to two electrodes with equal but opposite charges; the charge is just distributed over electrodes of more complicated shape. So we use the same definition for capacitance as in Eq. (8.4). The potential difference between any pair of electrodes of any of the capacitors is the same, equal to $(V_Q - V_{-Q})$. The total charge is simply the sum of the charges on the reference electrodes, i.e., $Q_{\text{tot}} = Q_1 + Q_2 + \cdots + Q_n$. From Eq. (8.4) it follows that the capacitance of such a connection of capacitors is

$$C_{\text{equiv}} = C_1 + C_2 + \cdots + C_n \quad (\text{F}). \quad (8.5)$$
$$\text{(Equivalent capacitance of a parallel connection of capacitors)}$$
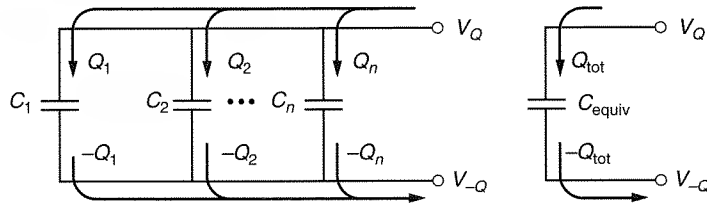


Figure 8.2 A parallel connection of capacitors
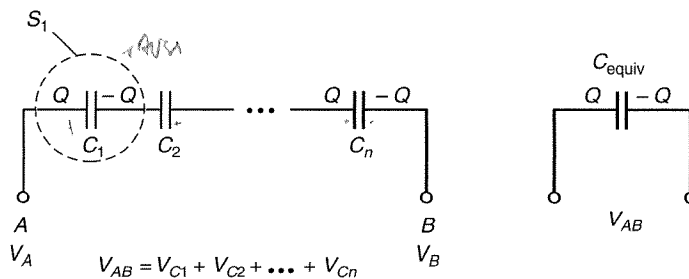
**Figure 8.3** A series connection of capacitors

**Example 8.3—Series connection of capacitors.** The series connection of capacitors, Fig. 8.3, is a little more complicated to analyze than the parallel case. It is now not obvious at all that such a structure is equivalent to the capacitor model of Fig. 8.1. This is indeed a different kind of structure. We have only two electrodes we can charge (the leftmost and the rightmost); the other electrodes are not accessible.

Note that the pairs of "internal" electrodes form conductive bodies with zero total charge. Assume that we charge the outermost electrodes with charges $Q$ and $-Q$. The left outer electrode will then induce a charge on the nearest electrode, and theoretically on all the others.

How large is this charge? Normally capacitors are made so that there is no field outside them if they are charged with equal but opposite charges. Assuming all the capacitors in Fig. 8.3 are of this type, and *only* in that case, if we enclose the first capacitor with a surface $S_1$ and apply Gauss' law, the total enclosed charge must be zero. This means that the induced charge on the second electrode from the left must be exactly $-Q$. This leaves $Q$ on the third electrode, which induces $-Q$ on the fourth one, etc. We see that *all the capacitors are charged with equal charge*, $Q$ and $-Q$.

Knowing the charge of all the capacitors, we know the voltage between their terminals:

$$V_{C1} = \frac{Q}{C_1}, \ V_{C2} = \frac{Q}{C_2}, \ \ldots , \ V_{Cn} = \frac{Q}{C_n}. \tag{8.6}$$

The total voltage, i.e., the voltage between the two outermost electrodes of the series connection, is the sum of these voltages. Since the charges corresponding to this voltage are $Q$ and $-Q$, the capacitance of this combined capacitor is given by

$$\frac{1}{C_{equiv}} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n} \quad \left(\frac{1}{F}\right). \tag{8.7}$$

*(Equivalent capacitance of a series connection of capacitors)*

Note that in a parallel connection of capacitors with greatly differing capacitances the dominant one is the one with the *largest* capacitance. If we have a series connection of such capacitors, the dominant one is the one with the *smallest* capacitance.

**Example 8.4—Parallel-plate capacitor filled with a homogeneous dielectric.** A parallel-plate capacitor consists of two parallel metal plates of areas $S$ charged with $Q_1 = Q$ and $Q_2 =$
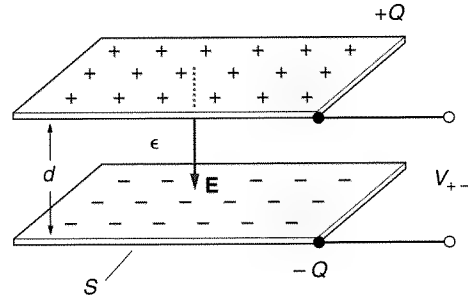
Figure 8.4 A parallel-plate capacitor

$-Q$ (Fig. 8.4). Assume that a homogeneous dielectric of permittivity $\epsilon$ is between the plates and that the distance between them, $d$, is small compared to the plate dimensions.

Under these assumptions, the field between the plates is very nearly the same as that between two uniformly charged planes. The electric displacement vector, $\mathbf{D}$, is normal to the plates. The surface charge density is $\sigma = Q/S$. Using the generalized Gauss' law we find that the intensity of the electric displacement vector $\mathbf{D} = \sigma = Q/S$. The electric field strength is hence $E = Q/(\epsilon S)$.

The voltage between the two plates corresponding to the given charge is now obtained easily. Since vector $\mathbf{E}$ between the plates is normal to them and constant,

$$V_Q - V_{-Q} = \int_+^- \mathbf{E} \cdot d\mathbf{l} = Ed = \frac{Qd}{\epsilon S},$$

so that

$$C = \epsilon \frac{S}{d} \quad \text{(F)}. \tag{8.8}$$

*(Capacitance of a parallel-plate capacitor)*

**Example 8.5—Parallel-plate capacitor with two dielectric layers.** Figure 8.5 shows a parallel-plate capacitor with two dielectrics, with the interface parallel to the plates. What is the capacitance in this case? The electric field is normal to the boundary between the two dielectrics, so we need to use the boundary condition for the displacement vector $\mathbf{D}$. Consequently, in dielectric 1, next to the left plate, $E_1 = \sigma/\epsilon_1 = Q/(\epsilon_1 S)$, and in the second dielectric, next to the right plate, $E_2 = Q/(\epsilon_2 S)$, where $S$ is the plate area. The vector $\mathbf{D}$ is normal to all the boundary surfaces. It is therefore continuous across the entire capacitor. The capacitance is given by

$$C = \frac{Q}{V_Q - V_{-Q}} = \frac{Q}{E_1 d_1 + E_2 d_2}.$$

(where $d_1$ and $d_2$ are thicknesses of the two layers (Fig. 8.5)). That is,

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2}, \tag{8.9}$$
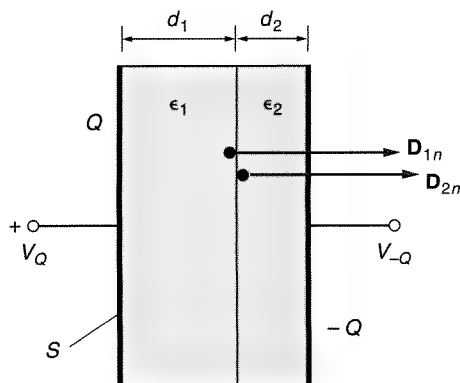
Figure 8.5 Parallel-plate capacitor with two different dielectrics between the plates

where $C_1$ and $C_2$ are the capacitances of parallel-plate capacitors with homogeneous dielectrics $\epsilon_1$ and $\epsilon_2$. This means that this capacitor looks like two capacitors in series.

**Example 8.6—Some other kinds of capacitors.** The expression $C = \epsilon S/d$ for the capacitance of the parallel-plate capacitor is often used even if the capacitor does not consist of two metal plates. For example, it is used for calculating the capacitance of the variable capacitor shown in Fig. 8.6a, where the capacitance is changed by turning one set of plates to overlap with the other set.

When variable capacitance is not needed and relatively large capacitance is required, capacitors like the one in Fig. 8.6b are used. Between two long ribbons made of aluminum foil is an insulating ribbon (for example, oily paper), and an insulating ribbon is also on the outside of one of the ribbons. The ribbons are tightly wrapped. The capacitance of such a capacitor can be precisely determined by the parallel-plate capacitor formula. (Note that because of the two insulating ribbons, the capacitance of the wrapped capacitor is *twice* that of the unwrapped capacitor.) Its capacitance can vary in a broad range, from about $10\,\text{pF}$ to about $100\,\mu\text{F}$.



(a)                                    (b)
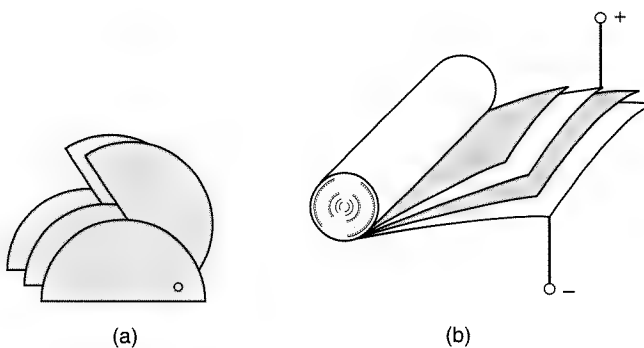
Figure 8.6 (a) A variable parallel-plate capacitor, and (b) a paper-insulator capacitor
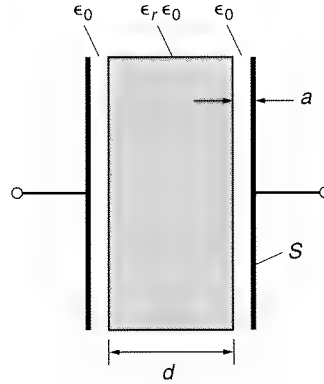
Figure 8.7 A parallel-plate capacitor with a
solid dielectric and a thin layer of air
between the dielectric and the electrodes

We see from the formula for the capacitance of the parallel-plate capacitor that very large capacitances can be obtained when $d$ is very small. In so-called electrolytic capacitors, the dielectric is a very thin layer of aluminum oxide (about $10^{-5}$ cm) deposited on the inside surface of a metal cap, and the capacitor is then filled with a conducting fluid. The fluid is one electrode, and the metal cap the other. In this way, the contact between the dielectric and the electrodes is very good. Electrolytic capacitors can have capacitances as large as hundreds and even thousands of microfarads.

In capacitors with solid insulators, the insulator and the metal electrodes might not have tight contact. In that case a thin layer of air is between them, as shown in Fig. 8.7. The capacitance of such a capacitor is

$$C = \frac{C_0 C_\epsilon}{C_0 + C_\epsilon}, \qquad C_0 = \epsilon_0 \frac{S}{2a}, \qquad C_\epsilon = \epsilon \frac{S}{d}. \tag{8.10}$$

Usually $a \ll d$ and $C_0 \gg C_\epsilon$, so the thin layer of air does not affect the capacitance too much. However, the electric field in the air layer is $E_0 = \epsilon_r E_d$, so it is larger than in the dielectric. Usually the dielectrics used in capacitors have a higher breakdown field than the 30 kV/cm for air. This air layer is a weak spot for high-voltage capacitors because breakdown would first occur in that layer.

**Example 8.7—Capacitance per unit length of a coaxial cable.** A coaxial cable, or coaxial transmission line, is used for guiding electromagnetic energy, especially at high frequencies. It consists of an inner wire conductor and an outer tubular conductor, coaxial with the wire; hence the cable name (Fig. 8.8). The coaxial cable is frequently nicknamed "coax."

Let the inner conductor have a radius $a$, and the outer conductor have an inside radius $b$. Usually the inner conductor is connected to the positive terminal of a voltage source, and the outer conductor is grounded. As a result, the inner conductor is charged along its length with $Q'$ coulombs/m (conditionally $Q' > 0$), and the outer conductor with $-Q'$ coulombs/m. Let the permittivity of the dielectric filling the cable be $\epsilon$.

Using Gauss' law on the surface $S_2$ for $b < r < c$, we find that all the charge on the outer conductor is distributed over its inside surface. The electric field is zero outside the coax. We
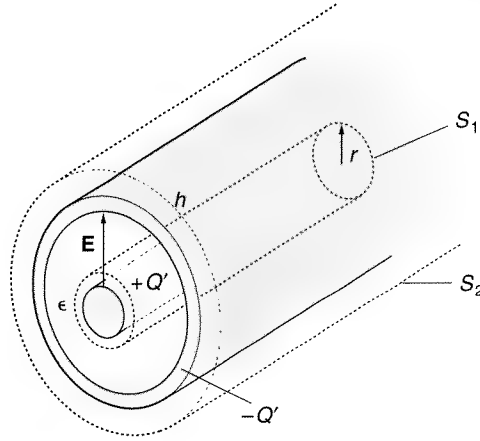
**Figure 8.8** A coaxial cable with a dielectric of permittivity $\epsilon$ between the two conductors. The inner conductor is charged positively, and the outer negatively (connected to ground).

want to find the capacitance per unit length of the coax. To do this, we need to find the electric field inside the cable, because $(V_{Q'} - V_{-Q'}) = \int \mathbf{E} \cdot d\mathbf{l}$ from the inner to the outer conductor.

For determining $\mathbf{E} - \mathbf{D}/\epsilon$, we use Gauss' law on the surface $S_1$, which is a cylinder of radius $r$ and height $h$. The flux through the cylinder bases is zero, so

$$\oint_{\text{cylinder}} \mathbf{D} \cdot d\mathbf{S} = D 2\pi r h = Q' h.$$

From here, we have

$$E = \frac{D}{\epsilon} = \frac{Q'}{2\pi \epsilon r}. \tag{8.11}$$

*(Electric field at a distance r from a long line charge)*

Note that the radius $a$ does not come into this expression, so it is valid for any radius of the inner conductor.

The voltage between the two conductors is now obtained as follows:

$$(V_{Q'} - V_{-Q'}) = \int_a^b E \, dr = \frac{Q'}{2\pi \epsilon} \int_a^b \frac{dr}{r} = \frac{Q'}{2\pi \epsilon} \ln \frac{b}{a}.$$

The capacitance per unit length of a coax is thus

$$C' = \frac{Q'}{V_+ - V_-} = \frac{2\pi \epsilon}{\ln(b/a)}. \tag{8.12}$$

*(Capacitance per unit length of a coax)*
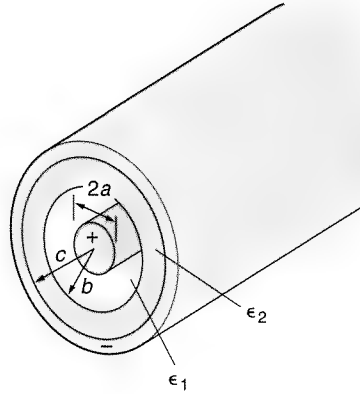
Figure 8.9 A high-voltage coaxial cable

We will see that all transmission lines are characterized by a so-called characteristic impedance. Coaxial cables are made most frequently with characteristic impedances of 50 Ω and 75 Ω. A typical value of the capacitance per unit length for a 50-Ω coaxial cable is about 100 pF/m, or 1 pF/cm.

As an example, a coaxial cable commonly used at high frequencies (e.g., in satellite receivers) is called RG-55/U and has the following parameters: $a = 0.5$ mm, $b = 2.95$ mm, $\epsilon_r = 2.25$. What is its capacitance per unit length?

**Example 8.8—Capacitance of a high-voltage coaxial cable.** In the expression for the electric field inside a coaxial cable, Eq. (8.11), we see that the electric field is the strongest right next to the inner conductor. In cables used for high-voltage applications, there is a danger of dielectric breakdown inside the coax. Therefore its inner conductor (where the field is the strongest) is frequently coated with a dielectric that has a high dielectric strength. The cross-section of a high-voltage coaxial cable is shown in Fig. 8.9. The electric field strength and displacement vectors are perpendicular to the boundary between the two dielectrics. Thus $D$ in the two dielectrics is given by the same expression, $D = Q'/(2\pi r)$, and $E$ in the two dielectrics is given by $E_1 = Q'/(2\pi \epsilon_1 r)$ and $E_2 = Q'/(2\pi \epsilon_2 r)$. The capacitance per unit length of this cable is

$$C' = \frac{Q'}{\int_a^b E_1 \, dr + \int_b^c E_2 \, dr} = \frac{C_1' C_2'}{C_1' + C_2'},$$

where $C_1' = 2\pi \epsilon_1 / \ln(b/a)$ and $C_2' = 2\pi \epsilon_2 / \ln(c/b)$.

**Example 8.9—Capacitance of a MOS capacitor.** The metal-oxide-semiconductor (MOS) capacitor is part of every metal oxide field-effect transistor (MOSFET), and millions of transistors are in every piece of electronic equipment. Figure 8.10 shows a MOS capacitor, which consists of a piece of $n$ semiconductor with a layer of dielectric (usually silicon dioxide) and a metal electrode deposited on top of the oxide. Similarly to a $pn$ diode, a depletion layer forms on the semiconductor side because the metal has many free electrons. The width of the depletion layer can be controlled by a voltage applied to the metal, and this effect is used in transistors, where the control electrode (gate) is essentially a MOS capacitor. However, due to

Figure 8.10 (a) A MOS capacitor, and (b) its equivalent series capacitor connection
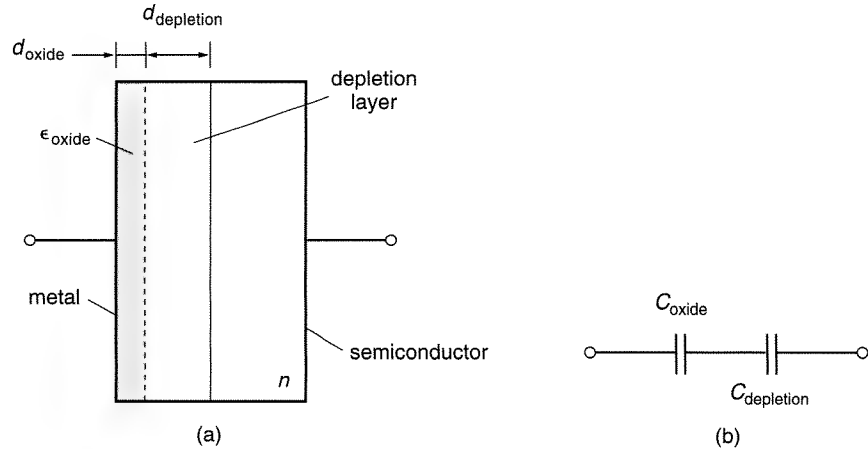
the presence of the oxide, the current flowing through the capacitor is essentially zero, and this makes the input impedance of a transistor very large.

The capacitance of the oxide is in series with the capacitance of the depletion layer, and the total capacitance is given by

$$\frac{1}{C_{\mathrm{MOS}}} = \frac{1}{C_{\mathrm{oxide}}} + \frac{1}{C_{\mathrm{depletion}}},$$

where $C_{\mathrm{oxide}} = \epsilon_{\mathrm{oxide}} S / d_{\mathrm{oxide}}$. We have seen in Examples 7.2 and 7.3 that the depletion layer is a uniform volume charge and that the electric field inside it is a linear function of the $x$ coordinate. Therefore, to find the capacitance we find the voltage by integrating the electric field from one end of the depletion layer to the other end. This is left as an exercise for the reader. Another useful exercise is to use superposition, as in Example 7.2, to find the electric field profile in a MOS capacitor.

*Questions and problems:*   Q8.1 to Q8.13, P8.1 to P8.23

## 8.3    Electrostatic Coupling in Multibody Systems

So far, we have considered an isolated conducting body and two bodies with equal but opposite charges. In practical applications we often have more than one or two conducting bodies. A multiconductor transmission line (bus) for connecting different parts of a computer is an example. We now consider this more general case, an electrostatic system consisting of an arbitrary number of charged conducting bodies.

We can adopt the reference point arbitrarily. To enable the analysis to apply to infinite structures as well (e.g., parallel, infinitely long wires), let us adopt as the reference *one of the conductors*. (We know that all points of a conductor in electrostatics are equipotential, and therefore we can adopt the entire body as the reference, not just
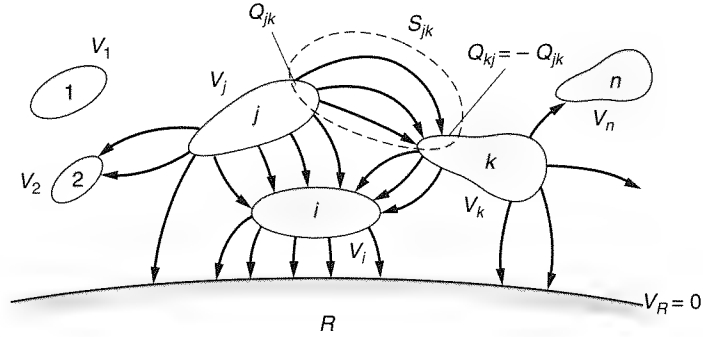
**Figure 8.11** A system of $n$ charged conducting bodies with the $(n+1)$th body, $R$, being the reference for potential

one of its points.) Most often, this reference body will be the earth or a convenient metallic part of the structure, such as the casing of an electronic device.

Consider a system of $n$ charged bodies with charges $Q_1, Q_2, \ldots, Q_n$, in addition to the reference body (Fig. 8.11). Let the dielectric be linear (but it need not be homogeneous). Is there a relationship between the charges on the bodies and their potentials, $V_1, V_2, \ldots, V_n$?

Because the system is linear, the principle of superposition applies. The potential of any of the bodies is obtained as

$$V_i = a_{i1}Q_1 + a_{i2}Q_2 + \cdots + a_{in}Q_n, \qquad i = 1, 2, \ldots, n. \tag{8.13}$$

*(Definition of coefficients of potential, $a_{ij}$)*

The coefficients $a_{ij}$ are termed, logically, the *coefficients of potential*. Note that their unit is 1/farad.

Provided that we know the coefficients $a_{ij}$, Eqs. (8.13) represent, in fact, a system of $n$ linear equations in $n$ unknowns. These unknowns can be the charges $Q_j$ of the bodies, but also their potentials, $V_i$. If the charges are known, Eqs. (8.13) represent the solution for the potentials. If the potentials are known, Eqs. (8.13) need to be solved for the charges, resulting in

$$Q_i = c_{i1}V_1 + c_{i2}V_2 + \cdots + c_{in}V_n, \qquad i = 1, 2, \ldots, n. \tag{8.14}$$

*(Definition of coefficients of electrostatic induction, $c_{ij}$)*

The coefficients $c_{ij}$ have several names. The most common is probably the *coefficients of electrostatic induction*. Their unit is the same as for capacitance, the farad.

Eqs. (8.14) can be rewritten in the following form:

$$Q_i = -c_{i1}(V_i - V_1) - c_{i2}(V_i - V_2) - \cdots + (c_{i1} + c_{i2} + \cdots + c_{in})V_i - \cdots$$
$$- c_{in}(V_i - V_n), \qquad i = 1, 2, \ldots, n. \tag{8.15}$$

Introducing new coefficients,

$$C_{ij} = -c_{ij} \text{ if } i \neq j, \text{ and } C_{ii} = c_{i1} + \cdots + c_{in}, \tag{8.16}$$

Eqs. (8.15) can be rewritten as

$$Q_i = C_{i1}(V_i - V_1) + C_{i2}(V_i - V_2) + \cdots + C_{ii}V_i + \cdots + C_{in}(V_i - V_n), \ i = 1, 2, \ldots, n. \quad (8.17)$$

*(Definition of coefficients of capacitance, $C_{ij}$)*

The coefficients $C_{ij}$ are known as the *coefficients of capacitance*. Their unit is also the farad.

If we know any of the three sets of coefficients, $a_{ij}$, $c_{ij}$, or $C_{ij}$, for a given system of conducting bodies, we can calculate mutual electrostatic effects in diverse circumstances. For example, we can assume that a body, instead of being at a desired potential, is unexpectedly grounded (at potential zero) and calculate the consequences of such an event. As another example, we can analyze the relative charge per unit length that one conductor of a multiconductor transmission line induces on the others, for given potentials of all the conductors. This, in fact, is an analysis of *electrostatic coupling*.

The only problem that needs to be solved is to determine in any way (analytically or experimentally) all the coefficients of one of the three sets, since they are derivable from each other. Let us explain, for example, how we can obtain the $a_{ij}$ coefficients. It is left as an exercise for the reader to imagine how to obtain in principle the coefficients $c_{ij}$ and $C_{ij}$.

Theoretically, we can find the coefficients $a_{ij}$ as follows. Assume that all the bodies except body $i$ are discharged, and that the charge on the $i$-th body is $Q_i$. Eqs. (8.13) then show that if we can measure the potentials $V_j$, $j = 1, 2, \ldots, n$ of the $n$ bodies, we can calculate $n$ potential coefficients $a_{ij}$. We repeat this procedure for all the $n$ bodies and obtain the complete set of the $a_{ij}$ coefficients.

**Example 8.10—Electrostatic coupling between a two-wire line and a parallel grounded wire.** Figure 8.12 shows a two-wire line at a height $d$ above ground. A wire at potential zero is parallel to the line, and all wires are in the same plane. Let conductors 1 and 2 of the line be charged with charges $Q'$ and $-Q'$, as indicated. Let the reference plane for potential be at the ground (Fig. 8.12). We wish to determine the charge per unit length induced on the grounded wire, $Q'_1$. It can be determined from the condition that the potential of the wire due to the three line charges, $Q'$, $-Q'$, and $Q'_1$, is zero.

The potential of a line charge is given by Eq. (6.9). Note that the distance of the three line charges from the reference plane for potential in this case is the same, equal to $d$. The poten-
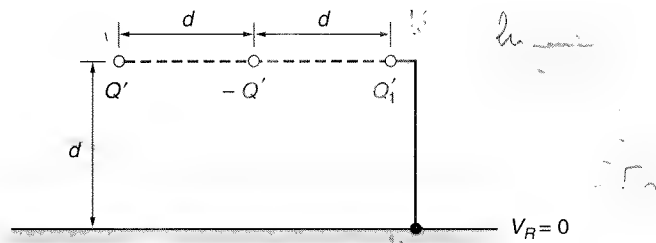


Figure 8.12 A two-wire line and a grounded wire running parallel to it

tial of the grounded wire is due to all three charges. Since it is zero, we obtain the following equation for the unknown charge per unit length $Q'_1$ of the grounded wire:

$$V_{\text{grounded wire}} = \frac{Q'}{2\pi\epsilon_0}\ln\frac{d/2}{2d} - \frac{Q'}{2\pi\epsilon_0}\ln\frac{d/2}{d} + \frac{Q'_1}{2\pi\epsilon_0}\ln\frac{3d/2}{a} = 0, \tag{8.18}$$

from which we easily find $Q'_1$.

If charges $Q'$ and $-Q'$ are time-varying, the induced charge on the wire will also be time-varying. This means that we would have an induced time-varying current in the wire due to electric coupling with the two-wire line. Such electric coupling is present in every multiconductor cable (such as a computer bus), between phone lines and power lines, and so on.

*Questions and problems:* Q8.14 to Q8.16, P8.24 to P8.27

## 8.4    Chapter Summary

1. Capacitance can be defined only under certain conditions. In the case of a single body, the condition is that it should be far from other bodies. For a two-body problem (the capacitor), the bodies should have equal but opposite charges, and the field of these charges should be restricted to the domain of the capacitor. Only in these circumstances are the familiar formulas for the capacitance of parallel and series connections of capacitors valid.

2. The ideal capacitor is an example of perfect electrostatic coupling between two bodies (by definition, there is no field outside the capacitor).

3. In a multibody system, mutual electrostatic coupling can be analyzed by means of any of three sets of coefficients, known as the coefficients of potential, coefficients of electrostatic induction, and coefficients of capacitance. These coefficients can (at least in principle) always be measured, but in many cases they can be calculated by numerical methods.

## QUESTIONS

**Q8.1.** A conducting body is situated in a vacuum. Prove that the potential of the body is proportional to its charge.

**Q8.2.** Repeat the preceding question if the body is situated in a linear (1) homogeneous or (2) inhomogeneous dielectric.

**Q8.3.** Two conducting bodies with charges $Q$ and $-Q$ are situated in a homogeneous linear dielectric. Prove that the potential difference between them is proportional to $Q$. Does the conclusion remain true if the dielectric is inhomogeneous (but still linear)?

**Q8.4.** Two conducting bodies with charges $Q$ and $-Q$ are situated in a homogeneous, but nonlinear, dielectric. Is the potential difference between them proportional to $Q$?

**Q8.5.** The capacitance of a diode is a function of the voltage between its terminals. Is this a linear or nonlinear capacitor?

**Q8.6.** Prove in your own words that a parallel connection of capacitors is indeed just a single unconventional capacitor.

**Q8.7.** Four metal spheres of radii $R$ are centered at corners of a square of side length $a = 3R$. Two pairs of the spheres are considered to be the electrodes of two capacitors, and are connected "in series." Is it possible to calculate the equivalent capacitance exactly using Eq. (8.7)? Explain.

**Q8.8.** A parallel-plate capacitor is connected to a source of voltage $V$. A dielectric slab is periodically introduced between the capacitor electrodes and taken out. Explain what happens with the capacitor charge.

**Q8.9.** Explain in your own words why the capacitance of a capacitor filled with a dielectric is larger than the capacitance of the same capacitor without the dielectric.

**Q8.10.** A negligibly thin metal foil is introduced between the plates of a parallel-plate capacitor, parallel to the plates. Is there any change in the capacitor capacitance? Can it be regarded as a series connection of two capacitors? Explain.

**Q8.11.** Repeat question Q8.10 assuming that the foil is not parallel to the plates.

**Q8.12.** Repeat question Q8.10 assuming the foil is thick.

**Q8.13.** A metal foil of thickness $a$ is introduced between and parallel to the plates of a parallel-plate capacitor that are a distance $d$ ($d > a$) apart. If the area of the foil and the capacitor plates is $S$, what is the capacitance of the capacitor without, and with, the foil?

**Q8.14.** Describe the procedure for measuring the coefficients of potential, $a_{ij}$, in Eq. (8.13).

**Q8.15.** Describe the procedure for measuring the coefficients of electrostatic induction, $c_{ij}$, in Eq. (8.14).

**Q8.16.** Describe the procedure for measuring the coefficients of capacitance, $C_{ij}$, in Eq. (8.17).

# PROBLEMS

**P8.1.** Two large parallel metal plates of areas $S$ are a distance $d$ apart, have equal charges of opposite sign, $Q$ and $-Q$, and the dielectric between the plates is homogeneous. Using Gauss' law, prove that the field between the plates is uniform. Calculate the capacitance of the capacitor per unit area of the plates.

**P8.2.** The permittivity between the plates of a parallel-plate capacitor varies as $\epsilon(x) = \epsilon_0(2 + x/d)$, where $x$ is the distance from one of the plates, and $d$ the distance between the plates. If the area of the plates is $S$, calculate the capacitance of the capacitor. Determine the volume and surface polarization charges if the plate at $x = 0$ is charged with a charge $Q$ ($Q > 0$), and the other with $-Q$.

**P8.3.** A parallel-plate capacitor with plates of area $S = 100\,\text{cm}^2$ has a two-layer dielectric, as in Fig. 8.5. One layer, of thickness $d_1 = 1\,\text{cm}$, has a relative permittivity $\epsilon_r = 3$, and a dielectric strength five times that of air. The other layer is air, of thickness $d_0 = 0.5\,\text{cm}$. How large a voltage will produce breakdown of the air layer, and how large does the voltage need to be to cause breakdown of the entire capacitor?

**P8.4.** A capacitor with an air dielectric was connected briefly to a source of voltage $V$. After the source was disconnected, the capacitor was filled with transformer oil. Evaluate the new voltage between the capacitor terminals.

**P8.5.** A capacitor of capacitance $C$, with a liquid dielectric of relative permittivity $\epsilon_r$, is connected to a source of voltage $V$. The source is then disconnected and the dielectric

drained from the capacitor. Determine the new voltage between the capacitor electrodes.

**P8.6.** Two conducting bodies with charges $Q$ and $-Q$ are situated in a linear, but inhomogeneous, dielectric. Prove that the potential difference between them is proportional to the charge $Q$.

**P8.7.** A parallel-plate capacitor has plates of area $S$ and a dielectric consisting of $n$ layers as in Fig. 8.5, with permittivities $\epsilon_1, \ldots, \epsilon_n$, and thicknesses $d_1, \ldots, d_n$. Evaluate the capacitance of the capacitor.

**P8.8.** Repeat problem P8.7 assuming that the layers are normal to the capacitor plates and that each layer takes the same amount of the capacitor plate area.

**P8.9.** Evaluate the maximal capacitance of the capacitor sketched in Fig. 8.6a if the plates are semicircular, of radius $R$, and the distance between adjacent plates is $d$. The dielectric is air.

**P8.10.** Evaluate the capacitance of the capacitor in Fig. 8.6b if the dielectric and aluminum ribbons are $a = 5\,\mathrm{cm}$ wide, $b = 2\,\mathrm{m}$ long, and $d = 0.1\,\mathrm{mm}$ thick. Assume the dielectric has a relative permittivity $\epsilon_r = 2.7$.

**P8.11.** Determine the polarization charges on all surfaces in Fig. 8.5.

**P8.12.** Determine the polarization charges on all dielectric surfaces in Fig. 8.9. Are there volume polarization charges anywhere? If so, where?

**P8.13.** One of two long, straight parallel wires is charged with a charge $Q'$ per unit length, and the other with $-Q'$. The wires have radii $a$ and are $d$ ($d \gg a$) apart. (1) Find the expression for the voltage between the wires and the capacitance per unit length of the line. Plot the magnitude of the electric field in a cross section of this two-wire line along the straight line joining the two wires. (2) At which points is it likely that the surrounding air will break down and ionize, given that a high-voltage generator is connected to the two wires? (3) If the wire radius is $a = 0.5\,\mathrm{mm}$, and the wires are $d = 1\,\mathrm{cm}$ apart, how large is the voltage of a voltage generator connected to the wires if the air at the wire surfaces breaks down?

**P8.14.** A spherical capacitor with two dielectrics is shown in Fig. P8.14. The inner radius is $a$, the outer radius is $b$, and the outer radius of the shell is $c$. The inner sphere is charged with $Q$ ($Q > 0$), and the outer shell with $-Q$. (1) Find the expression for the electric field everywhere and present your result graphically. (2) Find the expression for the
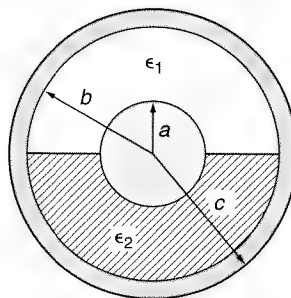


Figure P8.14 A two-dielectric spherical capacitor

capacitance of the capacitor. (3) If the outer shell is made to be much larger than the inner shell, what does the capacitance become and what does this mean physically?

**P8.15.** Two flat parallel conductive plates of surfaces $S = 0.05\,\text{m}^2$ are charged with $Q_1 = 5 \cdot 10^{-8}\,\text{C}$ and $Q_2 = -Q_1$. The distance between the plates is $D = 1\,\text{cm}$. Find the electric field strength vector at all points if a third, uncharged metal plate, $d = 5\,\text{mm}$ thick, is placed between the two plates $a = 2\,\text{mm}$ away from one of the charged plates and parallel to it. Plot the electric field strength before and after the third plate is inserted. Compare and explain. Find the capacitance between the charged plates without and with the third plate between them.

**P8.16.** The dielectric in a parallel-plate capacitor of plate area $S = 100\,\text{cm}^2$ consists of three parallel layers of relative permittivities $\epsilon_{1r} = 2$, $\epsilon_{2r} = 3$, and $\epsilon_{3r} = 4$. All three layers are $d = 1\,\text{mm}$ thick. The capacitor is connected to a voltage $V = 100\,\text{V}$. (1) Find the capacitance of the capacitor. (2) Find the magnitude of the vectors $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{P}$ in all dielectrics. (3) Find the free and polarization charge densities on all boundary surfaces.

**P8.17.** The surface area of each plate of a parallel-plate capacitor is $S = 100\,\text{cm}^2$, the distance between the plates is $d = 1\,\text{mm}$, and it is filled with a liquid dielectric of unknown permittivity. In order to measure the permittivity, we connect the capacitor to a source of voltage $V = 200\,\text{V}$. When the capacitor is connected to the source, it charges up, and the amount of charge is measured as $Q = 5.23 \cdot 10^{-8}\,\text{C}$ (the instrument that can measure this is called a ballistic galvanometer). Find the relative permittivity of the liquid dielectric.

**P8.18.** We wish to make a coaxial cable that has an electric field of constant magnitude. How does the relative permittivity of the dielectric inside the coaxial cable need to change as a function of radial distance in order to achieve this? The radius of the inner conductor is $a$ and the value of the relative permittivity right next to the inner conductor is $\epsilon_r(a)$. Find the capacitance per unit length of this cable.

**P8.19.** A capacitor in the form of rolled metal and insulator foils, Fig. 8.6b, needs to have a capacitance of $C = 10\,\text{nF}$. Aluminum and oily paper foils $a = 3\,\text{cm}$ wide are available. The thickness of the paper is $d = 0.05\,\text{mm}$, and its relative permittivity is $\epsilon_r = 3.5$. The thickness of the aluminum foil is also $0.05\,\text{mm}$. Find the needed length of the foil strips, as well as the maximum voltage to which such a capacitor can be connected. (Note that when rolled, the capacitance of the capacitor is twice that when the strips are not rolled.)

**P8.20.** A coaxial cable has two dielectric layers with relative permittivities $\epsilon_{1r} = 2.5$ and $\epsilon_{2r} = 4$. The inner conductor radius is $a = 5\,\text{mm}$, and the inner radius of the outer conductor is $b = 25\,\text{mm}$. (1) Find how the dielectrics need to be placed and how thick they need to be so that the maximum electric field strength will be the same in both layers. (2) What is the capacitance per unit length of the cable in this case? (3) What is the largest voltage that the cable can be connected to if the dielectrics have a breakdown field of $200\,\text{kV/cm}$?

**P8.21.** Figure P8.21 shows what is known as a capacitor bushing, which is used to insulate a high-potential conductor $A$ at its passage through the grounded wall $W$. The shaded surfaces represent thin dielectric sheets of permittivity $\epsilon$, and the thicker lines represent conducting foils placed between these sheets. Referring to Fig. P8.21, prove that the electric field intensity throughout the bushing is approximately the same, provided that $a_1 d_1 = a_2 d_2 = \ldots = a_4 d_4$.
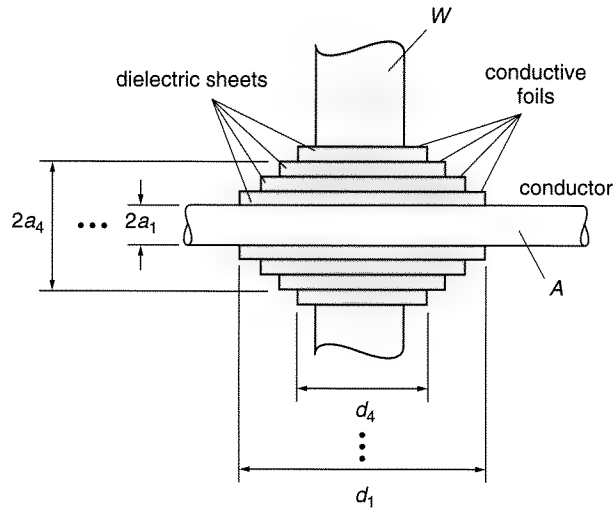
Figure P8.21 A capacitor bushing

*P8.22. Find the capacitance of a *pn* diode with a linear charge gradient, i.e., when the charge distribution on the *p* and *n* sides is as shown in Fig. P8.22. As in Example 7.2, you can assume that the charge on one side is much denser than that on the other side, and can therefore be assumed to be a charge sheet.
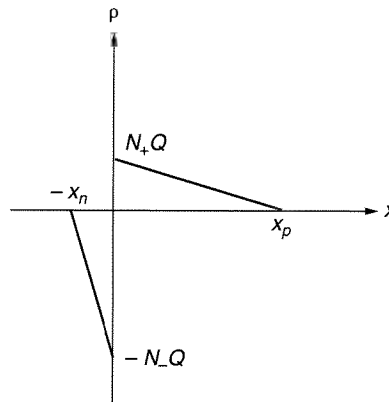


Figure P8.22 Linear charge profile in *pn* diode

P8.23. Plot the capacitance of a varactor diode as a function of the voltage across the diode. The capacitance of this diode is nonlinear and can be approximated with the following function of the voltage across the diode:

$$C_d(V) = \frac{C_0}{\sqrt{1 + (V/V_d)}},$$ (8.19)

where $C_0$ is the built-in capacitance (given) and $V_d$ is the built-in voltage of the diode (given). (This diode is used as an electrically variable capacitor because its capacitance can change significantly with applied voltage.)

**P8.24.** Find the expression for the capacitance $C_{12}$ between two bodies in terms of the coefficients of potential $a_{ij}$ defined by Eqs. (8.13). The two bodies have potentials $V_1$ and $V_2$, and the reference potential is the ground potential, as in Fig. P8.24.
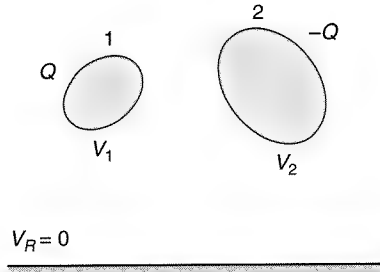


Figure P8.24 A capacitor above ground

**P8.25.** A two-wire line with charges $Q'$ and $-Q'$ runs parallel to the ground, with the two wires at different heights. The positively charged wire is $h_1$ above ground, and the negatively charged wire is $h_2$ above ground. The radii of both wires are $a$. Find the capacitance per unit length of such a line directly (from the definition of capacitance, and making use of images), and via the coefficients of potential defined by Eqs. (8.13), as follows:

  **(a)** Assuming the earth is at zero potential, that the left wire is charged with $Q'$, and that the other is uncharged, find the potential of both wires.

  **(b)** Repeat part a if the right wire is charged with $-Q'$ and the left wire is uncharged.

  **(c)** From the preceding and Eqs. (8.13), write down the expressions for the coefficients of potential $a_{ij}$ of the system.

  **(d)** Find the capacitance per unit length of the line.

**\*P8.26.** Prove that Eqs. (8.15) follow from Eqs. (8.14).

**\*P8.27.** Prove that from Eqs. (8.16) it follows that $c_{ij} = -C_{ij}$, and $c_{ii} = C_{i1} + \cdots + C_{in}$.

# 9

# Energy, Forces, and Pressure in the Electrostatic Field

## 9.1 Introduction

Measured by average human standards, electric energy, forces, and pressures are small. For example, it is virtually impossible to have an electric force of magnitude greater than a few newtons, or electric systems with energy exceeding a few thousand joules. Nevertheless, electric forces have surprisingly wide engineering applications. For example, purification of some ores, extraction of solid particles from smoke or dusty air, spreading of the toner in xerographic copying machines, and efficient and economical painting of car bodies are all based on electric forces.

Electrostatic energy is of equal engineering importance. For example, sufficient energy to destroy virtually any semiconductor device can easily be created in the field of a person charged by walking on a carpet. This is the meaning of the commonly used warning "static sensitive."

*Questions and problems:* Q9.1

## 9.2    Energy of a Charged Capacitor

In the preceding chapter, we defined *capacitance* and described and analyzed several types of capacitors. It is easy to understand that every charged capacitor contains a certain amount of energy. For example, the plates of a charged parallel-plate capacitor attract each other. If we let them move, they will perform a certain amount of work. In order for a system to do work, it must contain energy. Since the capacitor plates do not attract each other if they are not charged, it follows that some energy is stored in a *charged* capacitor. We can find how much energy there is by looking at what happens while a capacitor is being charged.

Consider a capacitor of capacitance $C$ that is initially not charged. We wish to charge its electrodes with $Q$ and $-Q$. To do this, we take small positive charges $dq$ from the negative electrode and take them over to the positive electrode. To move the charge against the electric forces ($dq$ is attracted by the negative electrode, and repelled by the positive electrode), we must do some work. Suppose that, at an instant during this process, the capacitor electrodes are charged with charges $q$ and $-q$ ($0 < q \le Q$). This means that the potential difference between them is $v = q/C$. By definition of the potential difference between the electrodes, the work we have to do against the electric forces in moving the next $dq$ from the negative to the positive electrode equals $dA = v\,dq = q\,dq/C$. So the total work that needs to be done to charge the capacitor electrodes with the desired charges, $Q$ and $-Q$, is

$$A = \int_0^Q \frac{q}{C}\,dq = \frac{Q^2}{2C} \qquad (\text{J}). \tag{9.1}$$

Since there were no losses in charging the capacitor, this work was transformed into potential energy of the capacitor. This energy we call the *electric energy*. Noting that $Q = CV$, the electric energy of a charged capacitor is thus given by the following equivalent expressions:

$$W_e = \frac{Q^2}{2C} = \frac{1}{2}QV = \frac{1}{2}CV^2 \qquad (\text{J}). \tag{9.2}$$

*(Energy of a charged capacitor)*

Let us look at a few examples. The largest possible energy of an air-filled parallel-plate capacitor with plate area $S = 1\,\text{dm}^2$ and with a distance between plates of $d = 1$ cm is

$$(W_e)_{\max} = \frac{1}{2}\epsilon_0 \frac{S}{d}E_{\max}^2 d^2 = \frac{1}{2}\epsilon_0 E_{\max}^2 Sd \simeq 4\,\text{mJ}, \tag{9.3}$$

since $E_{\max} \simeq 30\,\text{kV/cm}$. (The maximum energy corresponds to the maximum voltage, i.e., to the maximum electric field.) This is not very much energy from a human viewpoint (although it can destroy practically any semiconductor device).

If we consider a high-voltage capacitor, for example one where $V = 10\,\text{kV}$ and $C = 1\,\mu\text{F}$, we obtain instead

$$W_e = \frac{1}{2}CV^2 = 50\,\text{J}. \tag{9.4}$$

This is roughly equivalent to the potential energy of a 1-kg coconut that is 5 m above ground. The energy of high-voltage capacitors is clearly quite large, and touching their electrodes can be fatal.

*Questions and problems:*   Q9.2 to Q9.8, P9.1 to P9.4

## 9.3    Energy Density in the Electrostatic Field

The expression for the energy of a parallel-plate capacitor can be rewritten as

$$W_e = \frac{1}{2}CV^2 = \frac{1}{2}\epsilon\frac{S}{d}V^2 = \frac{1}{2}\epsilon E^2 Sd, \tag{9.5}$$

since $V/d = E$. The product $Sd$ is equal to the volume of the capacitor dielectric (i.e., the volume of the domain with the field). Therefore, no error will be made in computing the capacitor energy if we assume that it is distributed in the *entire* field, with a density

$$w_e = \frac{W_e}{v} = \frac{1}{2}\epsilon E^2. \tag{9.6}$$

*(Energy density, J/m³, in an electrostatic field)*

We will now show that this result is valid in general and not just for a parallel-plate capacitor. Let us look at a system of charged bodies in an arbitrary dielectric, as shown in Fig. 9.1. When we place thin aluminum foil exactly over an equipotential surface, we do not change the electric field. This is because we place a conducting surface, which must be equipotential, on an equipotential surface.

We can therefore place many thin aluminum foils on many equipotential surfaces, very close to each other, without changing the field. However, in this way we have divided up the space around the charged bodies into a very large number of small parallel-plate capacitors. The total energy of this system is given by the sum of all the little capacitor energies. The energy density of each of the capacitors is equal to $w_e = \epsilon E^2/2$, where $E$ is the electric field at that point, and $\epsilon$ the permittivity at that point. Consequently the energy of the whole system is given by

$$W_e = \int_v \frac{1}{2}\epsilon E^2 \mathrm{d}v. \tag{9.7}$$

*(Energy of an electric field)*
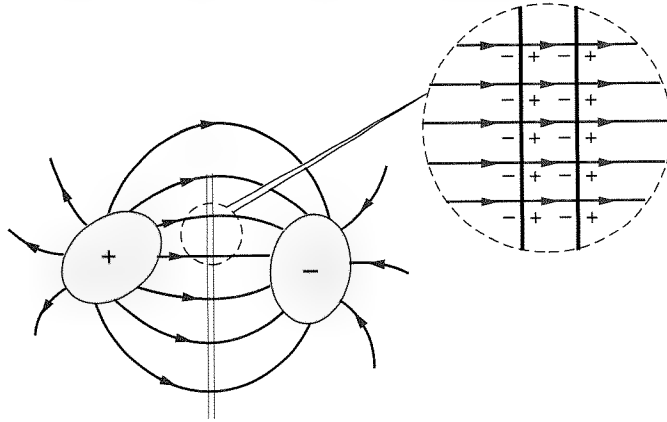
Figure 9.1 The electric field does not change when two aluminum foils are placed exactly at two close equipotential surfaces. Charges are induced on the surface of the foils (as shown in the enlarged circle), and the field between the foils is approximately uniform.

The integral in the equation is a volume integral *over the entire volume in which the electric field exists.*

**Example 9.1—Energy of a high-voltage coaxial cable.** In Example 8.8, we saw that a high-voltage coaxial cable consists of two dielectric layers and that the electric field in the two layers is given by

$$E_1 = \frac{Q'}{2\pi\epsilon_1 r} \qquad a < r < b$$

$$E_2 = \frac{Q'}{2\pi\epsilon_2 r} \qquad b < r < c.$$

The energy per unit length of the cable is the sum of the energies contained in the two dielectric layers:

$$W'_e = \int_{\text{layer1}} \frac{1}{2}\epsilon_1 E_1^2 dv' + \int_{\text{layer2}} \frac{1}{2}\epsilon_1 E_2^2 dv'.$$

Now $dv' = 2\pi r dr$, and $E_1$ and $E_2$ are given by the expressions at the beginning of the example. With respect to $r$, the first integral has limits from $a$ to $b$, and the second one from $b$ to $c$. After integrating, we get

$$W'_e = \frac{Q'^2}{2}\left[\frac{\ln(b/a)}{2\pi\epsilon_1} + \frac{\ln(c/b)}{2\pi\epsilon_2}\right].$$

If we use $W'_e = \frac{Q'^2}{2C}$, we get the same expression.

$$= (Q')^2/2C$$

We have concluded that energy contained in an electrostatic system can be determined if we assume that it is distributed throughout the field, with a density given in Eq. (9.6), even if the dielectric is a vacuum. In the case of dielectrics in the field, obviously at least some of the energy must be stored throughout the dielectric: to polarize the dielectric, the electric field needs to do some work *at the very point* where a dielectric molecule is, and *this molecule* acquires some energy. This means that the energy used to polarize a dielectric is distributed throughout the dielectric, just like the energy used in stretching a spring is distributed inside the entire spring.

In the case of a vacuum, however, such a physical explanation does not exist. How can we then state that the field in a vacuum also contains energy? In electrostatics, such a proof is not possible, but we shall see that in time-varying fields, the field *does* have energy distributed in a vacuum. For example, we know that a radio wave, which is but a combination of electric and magnetic fields, is able to carry a signal from the earth to Jupiter and back. This is a vast distance, and for a significant time the signal is neither on earth nor on Jupiter. It travels through a vacuum in between. It certainly carries some energy during this travel, because we are able to detect it.

*Questions and problems:*    Q9.9 to Q9.11, P9.5 to P9.12

## 9.4    Forces in Electrostatics

We started discussing electrostatics with Coulomb's law for the electric force between two point charges. Because the principle of superposition applies, it can be used as a basis for determining the electric force on any body in a system where we know the distribution of charges.

As an example, consider the two charged conducting bodies shown in Fig. 9.2, with a known surface charge distribution. Let us find the expression for the force $\mathbf{F}_{12}$ with which body 1 acts on body 2. To find this force, we divide body 2 into small patches $dS_2$ and determine the electric field strength $\mathbf{E}_1$ at all these patches due to the charge on body 1. The force is then obtained as

$$\mathbf{F}_{12} = \oint_{S_2} \sigma_2 \, dS_2 \mathbf{E}_1. \tag{9.8}$$
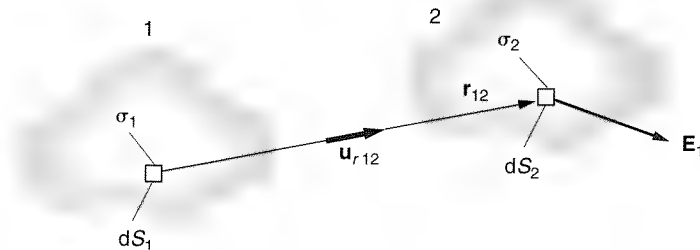


Figure 9.2 Finding the electric force between two large charged conducting bodies

In this equation, the field $\mathbf{E}_1$ is given by

$$\mathbf{E}_1 = \oint_{S_1} \frac{1}{4\pi\epsilon_0} \frac{\sigma_1 \, dS_1}{r_{12}^2} \mathbf{u}_{r12}, \tag{9.9}$$

where $\mathbf{r}_{12}$ is the vector directed from an element of body 1 toward an element of body 2, and $\mathbf{u}_{r12}$ is the unit vector along this direction.

**Example 9.2—Force between the plates of a parallel-plate capacitor.** Let us find the electric force that the electrodes of a parallel-plate capacitor of plate area $S$ exert on each other. We know that the charge is distributed practically uniformly on the electrode surface, i.e., the charge distribution is known. Let the capacitor be connected to a source of voltage $V$. The charge on the positive plate is then $Q = CV = \epsilon_0 S V/d$. We have found by Gauss' law that the electric field strength of the charge on the positive plate at the negative plate is $E_Q = Q/(2\epsilon_0 S)$. So using Eq. (9.8) we have

$$\mathbf{F}_{12} = \int_S (-\sigma) \, dS \, \mathbf{E}_Q = -Q\mathbf{E}_Q.$$

The force is attractive, as it should be, and its intensity is given by

$$F_{12} = QE_Q = \frac{Q^2}{2\epsilon_0 S} = \frac{1}{2}\epsilon_0 S \frac{V^2}{d^2}.$$

**Example 9.3—Magnitude of electric force in some typical devices.** What is the maximal electric force in a parallel-plate capacitor filled with air, with $S = 1 \, \mathrm{dm}^2$? The air breakdown field is $V/d \simeq 30 \, \mathrm{kV/cm}$, so we obtain $F_{12} \simeq 0.4 \, \mathrm{N}$, which amounts to the weight of about one quarter of a glass of water. Note that this is the *largest possible force*.

Another example is the force between the two wires of a two-wire line connected to a source of voltage $V$. The charge per unit length on the wires is $Q' = C'V = \pi\epsilon_0 V/\ln(d/a)$. At the place of the negatively charged wire, the positively charged wire produces a field $E_{Q'}$ equal to

$$E_{Q'} = \frac{Q'}{2\pi\epsilon_0 d}.$$

The force per unit length on the negatively charged wire is then

$$F'_{12} = -Q'E_{Q'} = -\frac{Q'^2}{2\pi\epsilon_0 d} = -\frac{\pi\epsilon_0 V^2}{2d[\ln(d/a)^2]}.$$

The minus sign tells us that the force is attractive, which it should be. Its maximal value for $a = 2.5 \, \mathrm{mm}$, $d = 1 \, \mathrm{m}$, and $E_{max} = 30 \, \mathrm{kV/cm}$ is $F'_{12} \simeq 0.00313 \, \mathrm{N/m}$. This is again quite a small force.

The two preceding examples illustrate the statement in the chapter introduction that in normal circumstances, electric forces acting between charged bodies are very

small. Therefore, they can be neglected most of the time. There are nevertheless many applications of the electrostatic forces, as will be discussed in Chapter 11.

*Questions and problems:*    Q9.12 to Q9.19, P9.13 to P9.16

## 9.5    Determination of Electrostatic Forces from Energy

We saw that we can find electric forces between charged bodies only if we know the charge distribution on them, which is rarely the case. Moreover, the previously discussed method cannot be used to determine forces on polarized bodies except in a few simple cases.

For example, suppose that a parallel-plate capacitor is partially dipped in a liquid dielectric, as in Fig. 9.3a. If the capacitor is charged, polarization charges exist only on the two vertical sides of the dielectric inside the capacitor. The electric force acting on them has only a horizontal component, if any. Yet experiment tells us that when we charge the capacitor, *there is a small but noticeable rise in the dielectric level between the plates*. How can we explain this phenomenon?

The answer lies in what happens not at the top of the dielectric but near the bottom edge of the capacitor. In that region, the dipoles in the dielectric orient themselves as shown in Fig. 9.3b. The net force on the dipoles points essentially upward and pushes the dielectric up between the plates. Although we can explain the nature of this force, based on what we have learned so far we have no idea how to calculate it. The method described next enables us to determine the electric forces in this and many other cases where the direct method fails. In addition, conceptually the same method is used for the more important determination of magnetic forces in practical applications.



(a)                                      (b)
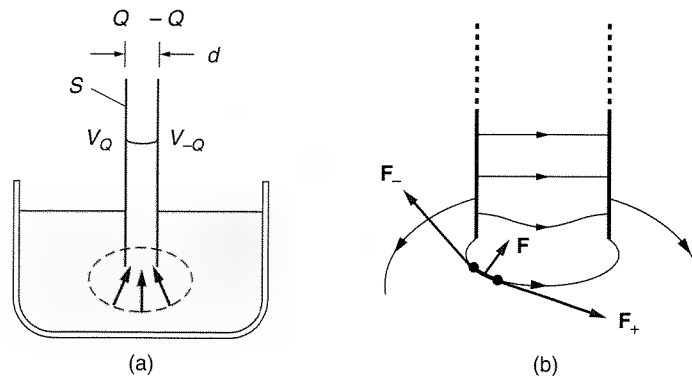
Figure 9.3 (a) When a parallel-plate capacitor dipped in a liquid dielectric is charged, the level between the plates rises due to electric forces acting on dipoles in the dielectric in the region around the edge of the capacitor, where the field is not uniform. (b) Enlarged domain of the capacitor fringing field in the dielectric, indicating the force on a dipole in a nonuniform field.
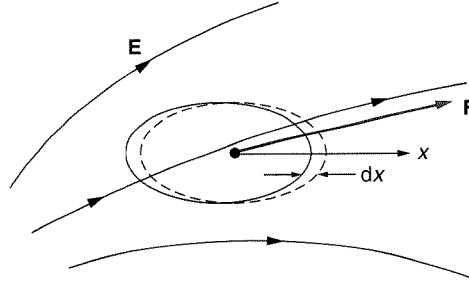
Figure 9.4 A body in an electrostatic system moved a small distance d$x$ by the electric force

Consider an arbitrary electrostatic system consisting of a number of charged conducting and polarized dielectric bodies. We know that there are forces acting on all these bodies. Let us concentrate on one of the bodies, for example the one in Fig. 9.4, that may be either a conductor or a dielectric. Let the *unknown* electric force on the body be **F**, as indicated in the figure.

Suppose we let the electric force move the body by a small distance d$x$ in the direction of the $x$ axis indicated in the figure. The electric force would in this case do work equal to

$$dA_{\text{el.force}} = F_x \, dx, \tag{9.10}$$

where $F_x$ is the projection of the force **F** on the $x$ axis.

At first glance we seem to have gained nothing by this discussion: we do not know the force **F**, so we do not know the work $dA_{\text{el.force}}$ either. However, we will now show that if we know how the electric energy of the system depends on the coordinate $x$, we can determine the work $dA_{\text{el.force}}$, and then from Eq. (9.10), the component $F_x$ of the force **F**. In this process, either (1) the charges on all the bodies of the system can remain unchanged or (2) the potentials of all the conducting bodies can remain unchanged.

Let us consider case (1) first. The charges can remain unchanged in spite of the change in the system geometry only if *none of the conducting bodies is connected to a source that could change its charge* (for example, a battery). Therefore, by conservation of energy, the work in moving the body can be done only at the expense of the electric energy contained in the system.

Let the system energy as a function of the coordinate $x$ of the body, $W_e(x)$, be known. The increment in energy after the displacement, $dW_e(x)$, is negative because some of the energy has been used for doing the work. Since work has to be a positive number, we have in this case $dA_{\text{el.force}} = -dW_e(x)$. Combining this expression with Eq. (9.10), the component $F_x$ of the electric force on the body is

$$F_x = -\frac{dW_e(x)}{dx} \quad \text{(charges kept constant)}. \tag{9.11}$$
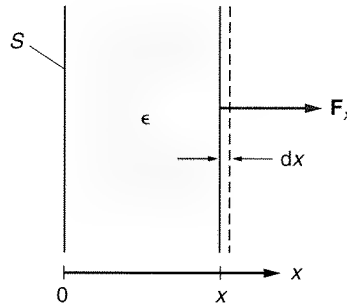
Figure 9.5 Determination of the force on the electrodes of a parallel-plate capacitor using Eq. (9.11)

**Example 9.4—Force acting on one plate of a parallel-plate capacitor.** In this example, we will find the electric force acting on one plate of a parallel-plate capacitor. The dielectric is homogeneous, of permittivity $\epsilon$, the area of the plates is $S$, and the distance between them is $x$. One plate is charged with $Q$ and the other with $-Q$ (Fig. 9.5). Let the electric force move the right plate by a small distance $dx$. The energy in the capacitor is given by $W_e(x) = Q^2/2C(x) = Q^2x/(2\epsilon S)$, so the force that tends to *increase* the distance between the plates is

$$F_x = -\frac{dW_e(x)}{dx} = -\frac{Q^2}{2\epsilon S}.$$

This is the same result as in Example 9.2, except for the sign. The minus sign tells us that the force tends to *decrease* the coordinate $x$, i.e., that it is attractive.

**Example 9.5—Force per unit length acting on a conductor of a two-wire line.** The wires of a two-wire line of radii $a$ are $x$ apart, and are charged with charges $Q'$ and $-Q'$. The energy per unit length of the line is

$$W_e'(x) = \frac{Q'^2}{2C'} = \frac{Q'^2}{2\pi\epsilon_0}\ln\frac{x}{a},$$

using $C'$ as calculated in problem P8.13. From Eq. (9.11) we obtain the force per unit length on the right conductor, tending to *increase* the distance between them, as

$$F_x = -\frac{dW_e}{dx} = -\frac{Q'^2}{2\pi\epsilon_0 x}.$$

This is the same as in Example 9.3, except for the minus sign. We know that this means only that the force tends to *decrease* the distance $x$ between the wires, i.e., that it is attractive.

**Example 9.6—Force acting on a dielectric partly inserted into a parallel-plate capacitor.** Let us find the electric force acting on the dielectric in Fig. 9.6. Equation (9.11) allows us to do this in a simple way. The capacitance of a capacitor such as this one is given by
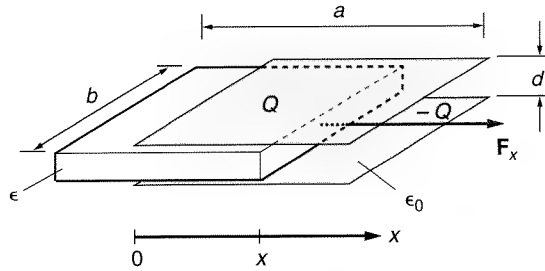
**Figure 9.6** Determination of the force on the dielectric partly inserted between the electrodes of a parallel-plate capacitor using Eq. (9.11)

$$C = C_1 + C_2 = \epsilon \frac{bx}{d} + \epsilon_0 \frac{b(a-x)}{d}$$

(see problem P8.8). The energy in the capacitor is

$$W_e(x) = \frac{Q^2}{2C} = \frac{Q^2}{2(C_1 + C_2)} = \frac{Q^2 d}{2b[\epsilon x + \epsilon_0(a-x)]}.$$

The derivative $dW_e(x)/dx$ in this case is a bit more complicated to calculate, and it is left as an exercise. The force is found to be

$$F_x = \frac{V^2}{2} \frac{b}{d} (\epsilon - \epsilon_0).$$

Note that this force is *always positive* because $\epsilon > \epsilon_0$. This means that the forces tend to pull the dielectric further in between the plates.

**Example 9.7—Rise of level of liquid dielectric partly filling a parallel-plate capacitor.** As a final example of the application of Eq. (9.11), let us determine the force that raises the level of the liquid dielectric between the plates of the capacitor in Fig. 9.3. Assume the dielectric is distilled water with $\epsilon_r = 81$, the width of the plates is $b$, their distance is $d = 1\,\text{cm}$, and the capacitor was charged by being connected to $V = 1000\,\text{V}$. The electric forces will raise the level of the water between the plates until the weight of the water between the plates becomes equal to this force. The weight is equal to

$$G = \rho_m xbdg,$$

where $\rho_m$ is the mass density of water and $g = 9.81\,\text{m/s}^2$. By equating this force to the force that we found in Example 9.6, we get

$$\rho_m xbdg = \frac{V^2}{2} \frac{b}{d} (\epsilon - \epsilon_0)$$

$$x = \frac{V^2}{2d^2 \rho_m g} (\epsilon - \epsilon_0) = 1.44\,\text{mm}.$$

So far, we have discussed examples of case (1), where the charges in a system were kept constant. Case (2) is finding forces from energy when the voltage, not the charge, of the $n$ conducting bodies of the system is kept constant (for example, we connect the system to a battery). When a body is moved by electric forces again by $dx$, some changes must occur in the charges on the conducting bodies, due to electrostatic induction. These changes are made at the expense of the energy in the sources (battery). So we would expect the energy contained in the electric field to increase in this case. It can be shown in a relatively straightforward way that the expression for the component $F_x$ of the electric force on the body in this case is

$$F_x = +\frac{dW_e(x)}{dx} \qquad \text{(potentials kept constant).} \tag{9.12}$$

Of course, this formula in all cases leads to the same result for the force as Eq. (9.11), but in some cases it is easier to calculate $dW_e/dx$ for constant potentials than for constant charges, and conversely.

**Example 9.8—Example 9.6 revisited.** Let us compute the force from Example 9.6 using Eq. (9.12) instead of Eq. (9.11), which we used in Example 9.6. Now we assume the potential of the two plates to be constant, and therefore express the system energy in the form

$$W_e(x) = \frac{1}{2}CV^2 = \frac{V^2}{2}\left[\epsilon \frac{bx}{d} + \epsilon_0 \frac{b(a-x)}{d}\right],$$

so that

$$F_x = +\frac{dW_e}{dx} = \frac{V^2}{2}\frac{b}{d}(\epsilon - \epsilon_0).$$

The result is easier to obtain than in Example 9.6.

*Questions and problems:*   P9.17 to P9.20

## 9.6    Electrostatic Pressure on Boundary Surfaces

In an electrostatic field there is pressure on all boundary surfaces. Although it is always small in terms of the pressure values we encounter around us (e.g., pressure of air in tires, pressure on pistons of combustible engines), it has interesting applications. Therefore we will derive the general expression for pressure on the boundary surface between two dielectrics, and estimate its magnitude.

Assume first that the boundary surface is tangential to the lines of the electric field strength vector (Fig. 9.7a). Let the electric forces push the surface by a small distance $dx$ from dielectric 2 toward dielectric 1, as in the figure. Since the lines of the vector **E** are tangential to the surface, the boundary conditions have not changed, so **E** remains the same. Therefore, the potential difference between any two bodies in the system remains the same as well. This means we have to use Eq. (9.12) to determine the force per unit area on the boundary surface.
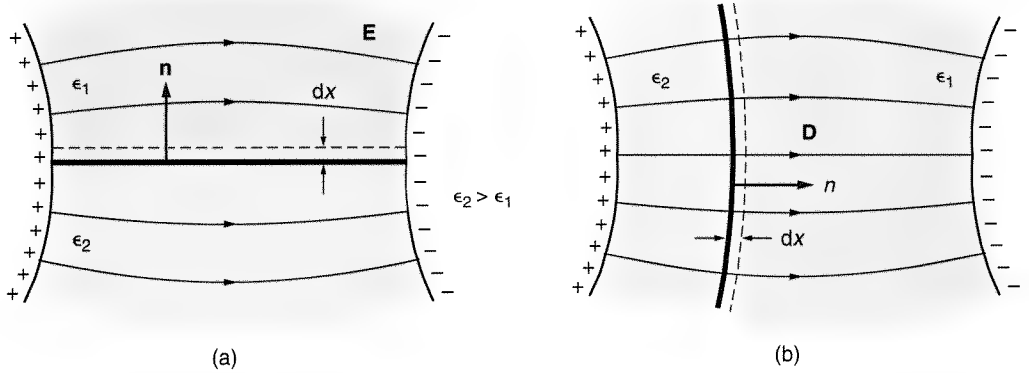
Figure 9.7 Boundary surfaces between two dielectrics. (a) The lines of the electric field strength **E** are tangential to the boundary. (b) The lines of the electric displacement vector **D** are normal to the boundary.

The energy in the system did change, since in the thin layer of thickness $dx$ the energy density before the displacement was $\epsilon_1 E_{tang}^2/2$, and after the displacement it became $\epsilon_2 E_{tang}^2/2$. If we consider a small patch of the boundary surface of area $\Delta S$, Eq. (9.12) yields

$$(F_x)_{\text{on } \Delta S} = +\frac{d}{dx}\left[\frac{1}{2}\left(\epsilon_2 E_{tang}^2 - \epsilon_1 E_{tang}^2\right) dx\, \Delta S\right],\tag{9.13}$$

from which the pressure on the boundary surface is

$$p_{Etang} = \frac{(F_x)_{\text{on } \Delta S}}{\Delta S} = \frac{1}{2}(\epsilon_2 - \epsilon_1)E_{tang}^2.\tag{9.14}$$

Note that the pressure acts *toward the dielectric of smaller permittivity*.

Consider now the case in Fig. 9.7b, where the boundary surface is such that the lines of the electric displacement vector **D** are normal to it. Assume again that due to electric forces, the surface is displaced by a small distance $dx$. The boundary conditions for vector **D** are satisfied, so it will not change. According to generalized Gauss' law, the charges on conducting bodies will therefore not be changed either. Hence this case corresponds to the formula in Eq. (9.11).

Again, in this case the energy density changed in the thin layer of thickness $dx$, so the force on a small patch of the boundary surface of area $\Delta S$ is found as

$$(F_x)_{\text{on } \Delta S} = -\frac{d}{dx}\left[\frac{1}{2}\left(\frac{D_{norm}^2}{\epsilon_2} - \frac{D_{norm}^2}{\epsilon_1}\right) dx\, \Delta S\right].\tag{9.15}$$

The electrostatic pressure in this case is thus

$$p_{Dnorm} = \frac{1}{2}\left(\frac{1}{\epsilon_1} - \frac{1}{\epsilon_2}\right)D_{norm}^2.\tag{9.16}$$

Note that in this case also the pressure acts *toward the dielectric of smaller permittivity*.

The lines of vector $\mathbf{E}$ are rarely tangential, and lines of vector $\mathbf{D}$ rarely normal, to boundary surfaces. When they are at an arbitrary angle with respect to the surface, the energy density in either of the two dielectrics can be expressed as

$$\frac{1}{2}\epsilon E^2 = \frac{1}{2}(\epsilon E_{\text{tang}}^2 + \epsilon E_{\text{norm}}^2) = \frac{1}{2}(\epsilon E_{\text{tang}}^2 + D_{\text{norm}}^2/\epsilon). \tag{9.17}$$

This means that the pressure due to the electrostatic field in the general case is given as the sum of the pressures in Eqs. (9.14) and (9.16):

$$p = \frac{1}{2}(\epsilon_2 - \epsilon_1)\left(E_{\text{tang}}^2 + \frac{D_{\text{norm}}^2}{\epsilon_1\epsilon_2}\right). \tag{9.18}$$

It is interesting that from Eq. (9.16) we can also obtain the pressure on the surface of a charged conductor. Let the conductor be medium 2, and assume that $\epsilon_2 \to \infty$, which implies that it is "infinitely polarizable," an electrostatic equivalent to a conductor. Replacing $\epsilon_1$ by $\epsilon$, Eq. (9.16) yields

$$p_{\text{on conductor surface}} = \frac{1}{2}\frac{D_{\text{norm}}^2}{\epsilon} = \frac{1}{2}\mathbf{E}\cdot\mathbf{D}. \tag{9.19}$$

The pressure is directed toward the dielectric.

**Example 9.9—Pressure on a liquid dielectric between plates of a parallel-plate capacitor.** Consider again the parallel-plate capacitor dipped into a liquid dielectric, Fig. 9.3a. Eq. (9.14) tells us immediately that there is an upward pressure on the upper surface of the dielectric. It is left as an exercise for the reader to show that the same result is obtained as before, but in a much simpler way.

**Example 9.10—Force acting on a plate of a parallel-plate capacitor.** The force on one of the plates of the parallel-plate capacitor (from Example 9.4) can now be obtained easily using Eq. (9.19). Note that we know the field on the plate surface if we know either the voltage between the plates or the charge of a plate (assumed to be distributed uniformly over it). The completion of this example is also left to the reader.

**Example 9.11—Magnitude of electrostatic pressure on a dielectric surface.** Let us now do a simple calculation that will tell us how strong electrostatic pressures can be. Imagine a slab of dielectric of $\epsilon_r = 4$ (say, quartz) is placed in an electric field perpendicular to the field lines (Fig. 9.8). Let us find the pressure on the front side of the slab for the strongest possible field in air, $E_0 = 30\,\text{kV/cm}$. Using Eq. (9.16), we obtain

$$p_{\text{Dnorm}} = \frac{1}{2\epsilon_0}\left(1 - \frac{1}{4}\right)D_0^2 = \frac{3}{8}E_0^2\epsilon_0 \simeq 30\,\text{Pa}.$$

In comparison, typical pressure inside a car tire is $200\,\text{kPa}$ (30 psi), or four orders of magnitude larger. [A pascal (Pa) is the SI unit for pressure equal to $\text{N/m}^2$. The psi stands for "pounds per square inch."]

*Questions and problems:*   Q9.20 to Q9.22, P9.21 and P9.22
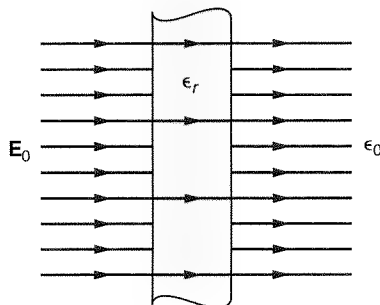
**Figure 9.8** A dielectric slab in an electric field. Electrostatic pressure on the slab side can be calculated using Eq. (9.16.)

## 9.7    Chapter Summary

1. Electrostatic energy, forces, and pressures are small when compared with the usual magnitude of these quantities around us. However, we will later show that they have considerable practical significance.

2. Electrostatic energy can be considered as a potential energy of a system of charges, or as distributed throughout the field with a density equal to $\frac{1}{2}\epsilon E^2$.

3. Electric forces can be obtained directly only if the charge distribution is known, which is rarely the case. Therefore a method for determining the forces based essentially on the law of conservation of energy has been derived. It enables the forces to be found from energy.

4. There is a pressure acting on all boundary surfaces in an electrostatic field. It is always directed toward the medium of lower permittivity.

<div align="center">

### QUESTIONS

</div>

**Q9.1.** What force drives electric charges that form electric current through circuit wires?

**Q9.2.** Capacitors of capacitances $C_1, C_2, \ldots, C_n$ are connected (1) in parallel, or (2) in series with a source of voltage $V$. Determine the energy in the capacitors in both cases.

**Q9.3.** A parallel-plate capacitor with an air dielectric, plate area $S$, and distance $d$ between plates is charged with a fixed charge $Q$. If the distance between the plates is increased by $dx$ ($dx > 0$), what is the change in electric energy stored in the capacitor? Explain the result.

**Q9.4.** Repeat question Q9.3 assuming that $dx < 0$.

**Q9.5.** A parallel-plate capacitor with an air dielectric and capacitance $C_0$ is charged with a charge $Q$. The space between the electrodes is then filled with a liquid dielectric of permittivity $\epsilon$. Determine the change in the electrostatic energy stored in the capacitor. Explain the result.

**Q9.6.** Can the density of electric energy be negative? Explain.

**Q9.7.** If you charge a 1-pF capacitor by connecting it to a source of 100 V, do you think the energy contained in the capacitor can damage a semiconductor device if discharged through it? Explain.

**Q9.8.** If you touch your two hands to the electrodes of a charged high-voltage capacitor, what do you think are the principal dangers to your body?

**Q9.9.** Explain in your own words why a polarized dielectric contains energy distributed throughout the dielectric.

**Q9.10.** Discuss whether a system of charged bodies can have zero total electric energy.

**Q9.11.** Can the electric energy of a system of charges be negative?

**Q9.12.** Explain in detail how you would calculate approximately the force $\mathbf{F}_{12}$ in Eq. (9.8), assuming that you know the charge distribution on the two bodies in Fig. 9.2.

**Q9.13.** If the field induces a dipole moment in a small body, it will also tend to move the body toward the region of stronger field. Sketch an inhomogeneous field and the dipole, and explain.

**Q9.14.** Under the influence of electric forces in a system, a body is rotated by a small angle. The system consists of charged, insulated conducting bodies. Is the energy of the system after the rotation the same as before, larger than before, or smaller than before? Explain.

**Q9.15.** If we say that $dW_e$ is negative, what does this mean?

**Q9.16.** Is weight a force? If it is, what kind of force? If it is not, what else might it be?

**Q9.17.** Is it possible to have a system of three point charges that are in equilibrium under the influence of their own mutual electric forces? If you can find such a system, is the equilibrium stable or unstable?

**Q9.18.** A soap bubble can be viewed as a small stretchable conducting ball. If charged, will it stretch or shrink? Do you think the change in size can be observed?

**Q9.19.** Explain why a charged body attracts *uncharged* small bodies of any kind.

**Q9.20.** Explain why, in Eq. (9.13), we subtracted the energy density in the first medium from the energy density in the second medium, and not the other way around.

**Q9.21.** A glass of water is introduced into an arbitrary inhomogeneous electric field. What is the direction of the pressure on the water surface?

**Q9.22.** Derive Eq. (9.19) from Eq. (9.18).

## PROBLEMS

**P9.1.** A bullet of mass 10 g is fired with a velocity of 800 m/s. How many high-voltage capacitors of capacitance 1 $\mu$F can you charge to a voltage of 10 kV with the energy of the bullet?

**P9.2.** A coaxial cable $h$ long, of inner radius $a$ and outer radius $b$, is first filled with a liquid dielectric of permittivity $\epsilon$. Then it is connected for a short time to a battery of voltage $V$. After the battery is disconnected, the dielectric is drained out of the cable. (1) Find the voltage between the cable conductors after the dielectric is drained out of the cable. (2) Find the energy in the cable before and after the dielectric is drained.

**P9.3.** A spherical capacitor with an air dielectric, of electrode radii $a = 10$ cm and $b = 20$ cm, is charged with a maximum charge for which there is still no air breakdown around the inner electrode of the capacitor. Determine the electric energy of the system.

**P9.4.** Repeat the preceding problem for a coaxial cable of length $d = 10\,\text{km}$, of conductor radii $a = 0.5\,\text{cm}$ and $b = 1.2\,\text{cm}$.

**P9.5.** Calculate the largest possible electric energy density in air. How does this energy density compare with a $0.5\,\text{J/cm}^3$ chemical energy density of a mixture of some fuel and compressed air?

**P9.6.** Show that half of the energy inside a coaxial cable with a homogeneous dielectric, of inner conductor radius $a$ and outer conductor radius $b$, is contained inside a cylinder of radius $a < r < \sqrt{ab}$.

**P9.7.** A metal ball of radius $a = 10\,\text{cm}$ is placed in distilled water ($\epsilon_r = 81$) and charged with $Q = 10^{-9}\,\text{C}$. Find the energy that was used up to charge the ball.

**P9.8.** A dielectric sphere of radius $a$ and permittivity $\epsilon$ is situated in a vacuum and is charged throughout its volume with volume density of free charges $\rho(r) = \rho_0 a/r$, where $r$ is the distance from the sphere center. Determine the electric energy of the sphere.

**P9.9.** Repeat the preceding problem if the volume density of free charges is constant, equal to $\rho$.

**P9.10.** Inside a hollow metal sphere, of inner radius $b$ and outer radius $c$, is a metal sphere of radius $a$. The centers of the two spheres coincide (concentric spheres), and the dielectric is air. If the inner sphere carries a charge $Q_1$ and the outer sphere a charge $Q_2$, what is the energy stored in the system?

**\*P9.11.** Prove *Thomson's theorem:* the distribution of static charges on conductors is such that the energy of the system of charged conductors is minimal.

**\*P9.12.** Prove that if an uncharged conductor, or a conductor at zero potential, is introduced into an electrostatic field produced by charges distributed on conducting bodies, the energy of the system decreases.

**P9.13.** An electric dipole of moment **p** is situated in a uniform electric field **E**. If the angle between the vectors **p** and **E** is $\alpha$, find the torque of the electric forces acting on the dipole. What do the electric forces tend to do?

**P9.14.** An electric dipole of moment $\mathbf{p} = Q\mathbf{d}$ is situated in an electric field of a negative point charge $Q_0$, at a distance $r \gg d$ from the point charge. If the vector **p** is oriented toward the point charge, find the total electric force acting on the dipole.

**P9.15.** A two-wire line has conductors with radii $a = 3\,\text{mm}$ and the wires are $d = 30\,\text{cm}$ apart. The wires are connected to a voltage generator such that the voltage between them is on the verge of initiating air ionization. (1) Find the electric energy per unit length of this line. (2) Find the force per unit length acting on each of the line wires.

**P9.16.** A conducting sphere of radius $a$ is cut into two halves, which are pressed together by a spring inside the sphere. The sphere is situated in air and is charged with a charge $Q$. Determine the force on the spring due to the charge on the sphere. In particular, if $a = 10\,\text{cm}$, determine the force corresponding to the maximal charge of the sphere in air for which there is no air breakdown on the sphere surface.

**P9.17.** Find the electric force acting on the dielectrics labeled 1 and 2 in the parallel-plate capacitor in Fig. P9.17. The capacitor plates are charged with $Q$ and $-Q$. Neglect edge effects.

**P9.18.** The inner conductor of the coaxial cable in Fig. P9.18 can slide along the cylindrical hole inside the dielectric filling. If the cable is connected to a voltage $V$, find the electric force acting on the inner conductor.

**P9.19.** One end of an air-filled coaxial cable with inner radius $a = 1.2\,\text{mm}$ and an outer radius of $b = 1.5\,\text{mm}$ is dipped into a liquid dielectric. The dielectric has a density of mass equal to $\rho_m = 0.8\,\text{g/cm}^3$, and an unknown permittivity. The cable is connected to a voltage
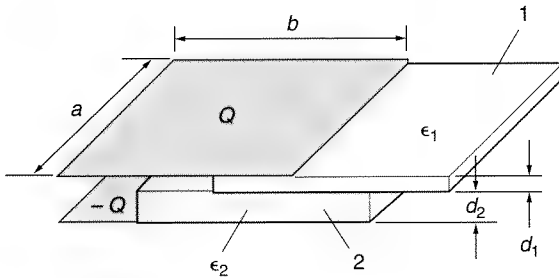
Figure P9.17 Three-dielectric capacitor
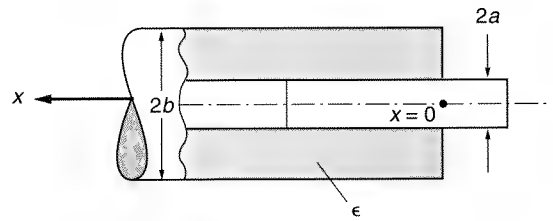


Figure P9.18 Coaxial cable with sliding conductor

$V = 1000$ V. Due to electric forces, the level of liquid dielectric in the cable is $h = 3.29$ cm higher than the level outside the cable. Find the approximate relative permittivity of the liquid dielectric, assuming the surface of the liquid in the cable is flat.

**P9.20.** The end of a coaxial cable is closed by a dielectric piston of permittivity $\epsilon$ and length $x$. The radii of the cable conductors are $a$ and $b$, and the dielectric in the other part of the cable is air. What is the magnitude and direction of the axial force acting on the dielectric piston, if the potential difference between the conductors is $V$?

**P9.21.** One branch of a U-shaped dielectric tube filled with a liquid dielectric of unknown permittivity is situated between the plates of a parallel-plate capacitor (Fig. P9.21). The voltage between the capacitor plates is $V$, and the distance between them $d$. The cross section of the U-tube is a very thin rectangle, with the larger side parallel to the electric field intensity vector in the charged capacitor. The dielectric in the tube above the liquid dielectric is air, and the mass density of the liquid dielectric is $\rho_m$. Assume that $h$ is the measured difference between the levels of the liquid dielectric in the two branches of the U-tube. Determine the permittivity of the dielectric.

**P9.22.** A soap bubble of radius $R = 2$ cm is charged with the maximal charge for which breakdown of air on its surface does not occur. Calculate the electrostatic pressure on the bubble.
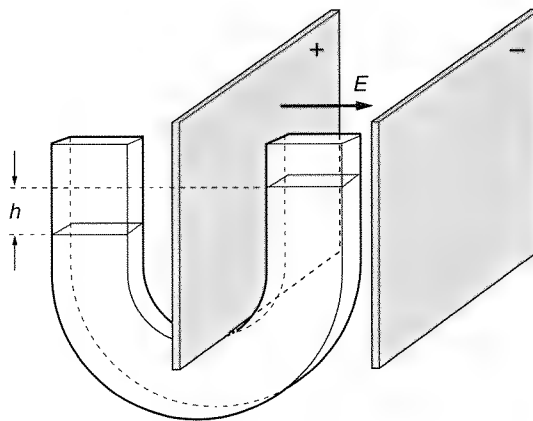


Figure P.9.21 Dielectric tube that is partially between the plates of a parallel-plate capacitor.

# 10

# Time-Invariant Electric Current in Solid and Liquid Conductors

## 10.1 Introduction

The term *time-invariant electric current* implies a steady, time-constant motion of a very large number of small charged particles. The term *current* is used because this motion is somewhat similar to the motion of a fluid. A typical example is the steady motion of free electrons inside a metallic conductor, but there are other types of time-invariant currents as well. What causes organized motion of large numbers of electrons (or other charges)? The answer is an electric field, which unlike in the electrostatic case, *does* exist inside current-carrying conductors. Time-invariant currents are frequently also called *direct currents*, abbreviated *dc*. A domain in which currents exist is known as the *current field*.

Inside a metallic conductor with no electric field present, a free electron (or any other type of free charge) moves chaotically in all directions, like a gas molecule. If there is an electric field inside the conductor, the electrons (negative charges) are accelerated in the direction opposite to that of the local vector **E**. This accelerated motion lasts until the electron collides with an atom. We can imagine that the electron

139
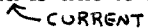
*DIRECTION TO

then stops, transfers the acquired kinetic energy to the atom, is again accelerated in the opposite *E, and so on. So the electrons acquire an average "drift" velocity under the influence of the field, and the result of this organized motion is an electric current.

There are three important consequences of this fact:

1. In solid and liquid conductors, where the average path between two collisions is very short, the drift velocity is in the direction of the force, i.e., *the charges follow the lines of vector* **E**.

2. Charges constantly lose the acquired kinetic energy to the atoms they collide with. This results in a more vigorous vibration of the atoms, i.e., a higher temperature of the conductor. This means that in the case of an electric current in conductors, the energy of the electric field is constantly converted into heat. This heat is known as *Joule's heat*. It is also frequently called *Joule's losses* because it represents a loss of electric energy.

3. In the steady time-invariant state, the motion of electric charges is time-invariant. The electric field driving the charges must in turn be time-invariant, and is therefore due to a time-constant distribution of charges. Such an electric field is *identical to the electrostatic field of charges distributed in the same manner*. This is a conclusion of extreme importance. All the concepts we derived for the electrostatic field (scalar potential, voltage, etc.) are valid for time-invariant currents.

Liquid conductors have pairs of positive and negative *ions*, which move in opposite directions under the influence of the electric field. The electric current in liquid conductors is therefore made of two streams of charged particles moving in opposite directions, but we have the same mechanism and the same effect of energy loss (Joule's heat) of current flow as in the case of a solid conductor. There is an additional effect, however, known as *electrolysis*—chemical changes in any liquid conductor that always accompany electric current.

In a class of materials called *semiconductors*, there are two types of charge carriers—negatively charged electrons and positively charged holes. In this case, the electric ~~field~~ is due to both types of charges and depends very much on their concentrations. ← CURRENT

In gases, electric current is also due to moving ions, but the average path between two collisions is much longer than for solid and liquid conductors. The mechanism of current flow is therefore quite different.

In solid and liquid conductors the number of charges taking part in an electric current is extremely large. To understand this, recall that a solid or liquid contains on the order of $10^{28}$ atoms per cubic meter. It is not easy to understand these huge numbers. Perhaps it would help if we consider a volume of about $(0.1 \text{ mm})^3$ (a cube 0.1 mm on each side), which is barely visible by the naked eye. This tiny volume contains about $10^{12}$ atoms, which is more than one hundred times the number of humans on our planet! It is evident from this example that the term "electric *current*" is indeed appropriate.

*Questions and problems:*   Q10.1

## 10.2   Current Density and Current Intensity: Point Form of Ohm's and Joule's Laws

Electric current in conductors is described by two quantities. The *current density vector*, **J**, describes the organized motion of charged particles at a point. The *current intensity* is a scalar that describes this motion in an integral manner, through a surface.

Let a conductor have $N$ free charges per unit volume, each carrying a charge $Q$ and having an average (drift) velocity **v** at a given point. The current density vector at this point is then defined as

$$\mathbf{J} = NQ\mathbf{v} \qquad \text{amperes per m}^2 \text{ (A/m}^2\text{).} \tag{10.1}$$

*(Definition of current density for one kind of charge carriers)*

Note that this definition implies that the current density vector of equal charges of opposite sign moving in opposite directions is the same. Of course, motion of different charges in opposite directions physically is different. However, experiments indicate that practically all effects (Joule's heat, chemical effects, magnetic effects) of an electric current depend on the product $Q\mathbf{v}$, so it is convenient to adopt this definition for the current density vector.

If there are several types of free charges inside a conductor, the current density is defined as a vector sum of the expression in Eq. (10.1). For example, let the current be due to the motion of free charge carriers of charges $Q_1, Q_2, \ldots, Q_n$, moving with drift velocities $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Let there be $N_1, N_2, \ldots, N_n$ of these charge carriers per unit volume, respectively. The current density vector is then given by

$$\mathbf{J} = \sum_{k=1}^{n} N_k Q_k \mathbf{v}_k \qquad \text{(A/m}^2\text{).} \tag{10.2}$$

*(Definition of current density for several kinds of charge carriers)*

The current intensity, $I$, through a surface is defined as the total amount of charge that flows through the surface during a small time interval, divided by this time interval. In counting this charge, opposite charges moving in opposite directions are added together. Thus

$$I = \frac{dQ_{\text{through } S \text{ in } dt}}{dt} \qquad \text{(C/s = A).} \tag{10.3}$$

*(Definition of current intensity through a surface)*

This can also be expressed in terms of the current density vector, as follows.

Consider a surface element $dS$ of the surface $S$ in Fig. 10.1. Let the drift velocity of charges at $dS$ be **v**, their charge $Q$, and their number per unit volume $N$. During the
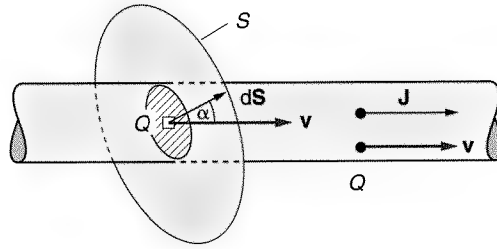
Figure 10.1 The current intensity $I$ through $S$ is equal to the flux of the current density vector $\mathbf{J}$ through $S$

time interval $dt$ the charges move by a distance $v\,dt$ in the direction of $\mathbf{v}$. Therefore, the charge that crosses $dS$ in $dt$ is

$$dQ_{\text{through } dS \text{ during } dt} = dS\,v\,dt\,\cos\alpha\,NQ, \tag{10.4}$$

where $\alpha$ is the angle between the velocity vector and the normal to the surface element. The total charge through $S$ during interval $dt$ is obtained as a sum (integral) of these elemental charges over the entire surface, and the current intensity is obtained by dividing this sum by $dt$. Noting that $dS\,v\,\cos\alpha NQ = J\,dS\,\cos\alpha = \mathbf{J}\cdot d\mathbf{S}$, we obtain

$$I = \int_S \mathbf{J}\cdot d\mathbf{S} \quad \text{(A)}. \tag{10.5}$$

(Definition of current intensity through a surface in terms of the current density vector)

The unit for current is an *ampere* (A), equal to a coulomb per second (C/s). The unit for current density is $A/m^2$.

We now know that electric current in a conductor is produced by an electric field. We also know that in solid and liquid conductors the vectors $\mathbf{J}$ and $\mathbf{E}$ are in the same direction. For most conductors, vector $\mathbf{J}$ is a linear function of $\mathbf{E}$,

$$\mathbf{J} = \sigma\mathbf{E} \quad [\sigma - \text{siemens per meter (S/m)}]. \tag{10.6}$$

*[Point (local) form of Ohm's law]*

Conductors for which (10.6) is valid are called *linear conductors*. The constant $\sigma$ is known as the *conductivity* of the conductor. The unit for conductivity is *siemens per meter* (S/m).

The reciprocal value of $\sigma$ is designated by $\rho$ and is known as the *resistivity*. The unit for resistivity is *ohm · meter* ($\Omega \cdot m$). Equation (10.6) can be written in the form

$$\mathbf{E} = \rho\mathbf{J} \qquad [\rho - \text{ohm} \cdot \text{meter} \, (\Omega \cdot \text{m})]. \tag{10.7}$$

[Point (local) form of Ohm's law]

Both Eqs. (10.6) and (10.7) are known as the *point form of Ohm's law* for linear conductors because they give a relationship between the two field quantities at every point inside a conductor.

For metallic conductors, conductivities range from about 10 MS/m (iron) to about 60 MS/m (silver). The conductivity of seawater is about 4 S/m, that of ground (soil) is between $10^{-2}$ and $10^{-4}$ S/m, and conductivities of good insulators are less than about $10^{-12}$ S/m.

We have already explained from a physical standpoint that there is a permanent transformation of electric energy into heat in every current field. Let us now derive the expression, known as *Joule's law in point form*, for the volume density of power in this energy transformation.

Let there be $N$ charge carriers $Q$ in the conductor, and let their local drift velocity be $\mathbf{v}$. The electric force on each charge is $Q\mathbf{E}$. The work done by the force when moving the charge during a time interval $dt$ is equal to $Q\mathbf{E} \cdot (\mathbf{v}\,dt)$. The work done in moving all the $N\,dv$ charges inside a small volume $dv$ is therefore

$$dA_{\text{el.forces}} = Q\mathbf{E} \cdot (\mathbf{v}\,dt)N\,dv = \mathbf{J} \cdot \mathbf{E}\,dv\,dt \qquad (\text{J}). \tag{10.8}$$

If we divide this expression by $dv\,dt$, we get the desired power per unit volume (volume power density)—the electric power that is lost to heat:

$$p_J = \frac{dP_J}{dv} = \mathbf{J} \cdot \mathbf{E} = \frac{J^2}{\sigma} = \sigma E^2 \quad \text{watts/m}^3 \, (\text{W/m}^3). \tag{10.9}$$

(Joule's law in point form)

If we wish to determine the power of Joule's losses in a domain of space, we just have to integrate the expression in Eq. (10.9) over that domain:

$$P_J = \int_v \mathbf{J} \cdot \mathbf{E}\,dv \quad \text{watts (W)}. \tag{10.10}$$

(Joule's losses in a domain of space)

**Example 10.1—Fuses.** Electrical devices are frequently protected from excessive currents by fuses, one type of which is sketched in Fig. 10.2. The fuse conductor is made to be much thinner than the circuit conductors elsewhere. For example, let the radius of the circuit conductor be $n$ times that of the fuse. If a current of intensity $I$ exists in the circuit, the volume density of Joule's losses in the thin conductor section is $n^4$ larger than those in the other section. In the case of excessive current, therefore, the thin conductor section melts long before the normal section is heated up. When the fuse melts, it becomes an open circuit and does not allow any
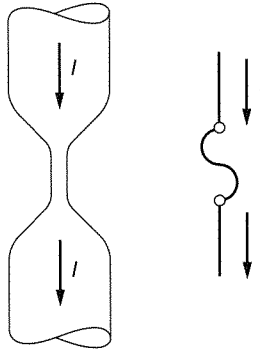
Figure 10.2 A simple model of a fuse

further current to flow and possibly damage the device protected by the fuse. Usually the thin conductor is a metal with a conductivity smaller than that of the thick wire.

*Questions and problems:*   Q10.2 and Q10.3, P10.1 to P10.7

## 10.3   Current-Continuity Equation and Kirchhoff's Current Law

Experiments tell us that electric charge cannot be created or destroyed. This is known as the *law of conservation of electric charge*. The continuity equation is the mathematical expression of this law. Its general form is valid for time-varying currents, but it can easily be specialized for time-invariant currents.

Consider a closed surface $S$ in a current field. Let **J** be the current density (a function of coordinates and, in the general case considered here, of time). The definition of current intensity applies to any surface, so it applies to a closed surface as well. The current intensity, $i(t)$, through $S$, with respect to the outward normal, is given by Eq. (10.3):

$$i(t) = \frac{dq(t)_{\text{out of } S \text{ in } dt}}{dt}. \tag{10.11}$$

According to the law of conservation of electric charge, if some amount of charge leaves a closed surface, the charge of opposite sign inside the surface must increase by the same amount. So we can write Eq. (10.11) as

$$i(t) = -\frac{dq(t)_{\text{inside } S \text{ in } dt}}{dt}. \tag{10.12}$$

The current intensity can also be written in the form of Eq. (10.5), and

$$\frac{dq(t)_{\text{inside } S \text{ in } dt}}{dt} = \frac{d}{dt} \int_v \rho(t) dv, \tag{10.13}$$

where $v$ is the volume enclosed by $S$. Recall that by convention we always adopt the outward unit vector normal to a closed surface. Thus Eq. (10.12) can be rewritten in

the form

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = -\frac{d}{dt} \int_v \rho(t) dv. \tag{10.14}$$

*(General form of the current continuity equation, where surface S may vary in time)*

This is the *current continuity equation*. Note that we can imagine the surface $S$ to change in time, in which case the form of the continuity equation in Eq. (10.14) must be used. This, however, is needed only in rare instances. If the surface $S$ does not change in time, the time derivative acts on $\rho$ only. Because $\rho$ is a function of both time and space coordinates, the ordinary derivative needs to be replaced by a partial derivative, and we obtain a much more important form of the current continuity equation:

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = -\int_v \frac{\partial \rho(t)}{\partial t} dv. \tag{10.15}$$

*(Current continuity equation for a time-invariant surface)*

Although the current continuity equation is not a field equation, it is of fundamental importance in the analysis of electromagnetic fields because only sources (charges and currents) satisfying this equation can be real sources of the field.

Now let the current field be constant in time, in which case the charge density is also constant in time. The partial derivative of $\rho$ on the right side of Eq. (10.15) is then zero, and both forms of the current continuity equation reduce to

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = 0. \tag{10.16}$$

*(Generalized Kirchhoff's current law)*

This equation tells us that in time-constant current fields the amount of charge that flows into a closed surface is *exactly* the same as that which flows out of it. Equation (10.16) represents, in fact, the generalized form of the familiar *Kirchhoff's current law* from circuit theory. Indeed, if a surface $S$ encloses a node of a circuit, there are currents only through the circuit branches, and Eq. (10.16) becomes

$$\sum_{k=1}^{n} I_k = 0. \tag{10.17}$$

We know that Kirchhoff's current law in this form is applied also to circuits with time-varying currents. Considering the preceding discussion, it should be clear that in such cases it is only approximate. (Can you explain why?)
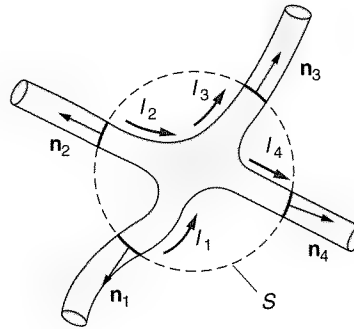
Figure 10.3 Application of generalized Kirchhoff's current law to a node with four wires

**Example 10.2—Continuity equation applied to a circuit node.** Let the surface $S$ enclose a node with four wires (Fig. 10.3), with dc currents $I_1$, $I_2$, $I_3$, and $I_4$. The vector $\mathbf{J}$ is nonzero only over small areas of $S$ where the wires go through the surface. There, the flux of $\mathbf{J}$ is simply the current intensity in that wire, so that Eq. (10.16) yields $-I_1 - I_2 + I_3 + I_4 = 0$, which is what we would get if we simply applied Kirchhoff's current law to the node. How are the signs of the currents determined and what do they correspond to in Eq. (10.16)?

*Questions and problems:*    Q10.4 and Q10.5

## 10.4   Resistors: Ohm's and Joule's Laws

A resistor is a resistive body with two equipotential contacts. A resistor of general shape is shown in Fig. 10.4. Assume that the material of the resistor is linear. We know that the resistivity, $\rho$, for linear materials does not depend on the current density. Then the current density, $\mathbf{J}$, is proportional at all points of the resistor to the current intensity $I$ through its terminals. Therefore, the electric field vector in the
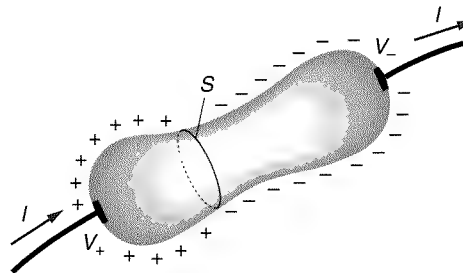


Figure 10.4 A resistor consists of a resistive body with two metallic (equipotential) contacts (the resistor terminals)

resistor material, $\mathbf{E} = \rho\mathbf{J}$, and the potential difference between its terminals are also proportional to the current intensity,

$$V_+ - V_- = RI \qquad [R - \text{ohms } (\Omega)], \tag{10.18}$$

where $R$ is a constant. This equation is known as *Ohm's law*. Resistors for which this equation holds are called *linear resistors*. The constant $R$ is called the *resistance* of the resistor. In some instances it can be computed starting from the defining formula in Eq. (10.18), but it can always be measured. The unit for resistance is the *ohm* ($\Omega$).

The reciprocal of resistance is called the *conductance*, $G$. Its unit is called the *siemens* (S). In the United States, sometimes the mho (ohm backwards) is used instead, but this is not a legal SI unit and we will not use it.

**Example 10.3—Resistance of a straight wire segment.** As an example of calculating resistance, consider a straight wire of resistivity $\rho$, length $l$, and cross-sectional area $S$. Let the current intensity in the wire be $I$. The current density vector is parallel to the wire axis, and its magnitude is $J = I/S$. The electric field vector is therefore also parallel to the wire axis, and its magnitude is $E = \rho J = \rho I/S$. The potential difference between the ends of the wire segment is $V_1 - V_2 = El = \rho l I/S$. So the resistance of the wire segment is

$$R = \rho\frac{l}{S} \qquad (\Omega). \tag{10.19}$$

Consider now a resistor of resistance $R$. Let the current intensity in the resistor be $I$, and the voltage between its terminals $V$. During a time interval $t$, a charge equal to $Q = It$ flows through the resistor. This charge is transported by electric forces from one end of the resistor to the other end. From the definition of voltage, the work done by electric forces is

$$A_{\text{el.forces}} = QV = VIt \qquad (\text{J}). \tag{10.20}$$

Because of energy conservation, an energy equal in magnitude to this work is transformed into heat inside the resistor:

$$W = VIt = RI^2t = \frac{V^2}{R}t \qquad (\text{J}). \tag{10.21}$$

Since the process of transformation of electric energy into heat is constant in time, the power of this transformation of energy is $W/t$, that is,

$$P = VI = RI^2 = \frac{V^2}{R} \qquad (\text{W}). \tag{10.22}$$

*(Joule's law)*

This is the familiar *Joule's law* from circuit theory. It is named after the British physicist James Prescott Joule (1818–1889), who established this law experimentally.

*Questions and problems:*   Q10.6 to Q10.9, P10.8 to P10.14

## 10.5  Electric Generators

We know that actual sources of the electric field are electric charges. We also know that we must remove some charges from a body, or put them on a body, in order to obtain excess electric charges on it. This obviously cannot be done by the electric forces themselves. Devices that do this must use some nonelectric energy to separate one type of charge from another. Such devices are known as *electric generators*.

An electric generator is sketched in Fig. 10.5. In a region of the generator there are nonelectric forces, known as *impressed forces*, that separate charges of different signs on the two generator terminals, or electrodes, denoted in the figure by "−" (negative) and "+" (positive). These forces can be diverse in nature. For example, in chemical batteries these are chemical forces; in thermocouples these are forces due to different mobilities of charge carriers in the two conductors that make the connection; in large rotating generators these are magnetic forces acting on charges inside conductors.

Impressed forces, by definition, act only on charges. Therefore they can always be represented as a product of the charge on which they act and a vector quantity that must have the dimension of the electric field strength:

$$\mathbf{F}_{\text{impressed}} = Q\mathbf{E}_i. \tag{10.23}$$

The vector $\mathbf{E}_i$ is not necessarily an electric field strength (although it does have the same dimension and unit). It is known as the *impressed electric field strength*. The region of space it exists in is known as the *impressed electric field*. The impressed electric field is a concept used often in electromagnetic theory. For example, in the analysis of radar wave scattering from a radar target, the radar wave is considered to be a field impressed on the target.

The *electromotive force*, or *emf*, of a generator is defined as the work done by impressed forces in taking a unit charge through the generator, from its negative to
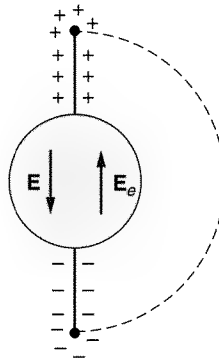


Figure 10.5 Sketch of an electric generator

its positive terminal:

$$\mathcal{E} = \left\{ \int_{-}^{+} \mathbf{E}_i \cdot d\mathbf{l} \right\}_{\text{through the generator}} \qquad \text{(definition of emf).} \qquad (10.24)$$

By simple reasoning, the emf of a generator can be expressed in terms of the voltage between open-circuited generator terminals, as follows. Assume that the generator is open-circuited (no current is flowing through the generator). Then at all its points, the electric field strength due to separated charges and the impressed electric field strength must have equal magnitudes and opposite directions, i.e., $\mathbf{E} = -\mathbf{E}_i$ (otherwise free charges in the conducting material of the generator would move). If we substitute this into Eq. (10.24) and exchange the integration limits, we get

$$\mathcal{E} = \left\{ \int_{+}^{-} \mathbf{E} \cdot d\mathbf{l} \right\}_{\text{any path}} = V_+ - V_-. \qquad (10.25)$$

The electric field strength $\mathbf{E}$ is due to time-constant charges, i.e., it is an electrostatic field. The line integral can therefore be taken along *any* path, including one outside the generator (dashed line in Fig. 10.5). This means that we can measure the emf of a generator by a voltmeter with the voltmeter leads connected in any way to the terminals of an open-circuited generator. We will see that this is not the case for time-varying voltages.

Because the generator is always made of a material with nonzero resistivity, some energy is transformed into heat in the generator itself. Therefore, we describe a generator by its internal resistance also. This is simply the resistance of the generator in the absence of the impressed field.

*Questions and problems:*   Q10.10 to Q10.12, P10.15 and P10.16

## 10.6   Boundary Conditions for Time-Invariant Currents

Current fields often exist in media of different conductivities separated by boundary surfaces. We now formulate boundary conditions for this case.

Consider the boundary surface sketched in Fig. 10.6. If we apply the continuity equation for time-invariant currents to the small, coinlike closed surface, we immediately see that the normal components of the current density vector in the two media must be the same:

$$J_{1n} = J_{2n}, \qquad \text{or} \qquad \sigma_1 E_{1n} = \sigma_2 E_{2n}. \qquad (10.26)$$

We know that the electric field in a current field has the same properties as the electrostatic field. Therefore the same boundary condition for the tangential component of the vector $\mathbf{E}$ on a boundary applies:

$$E_{1t} = E_{2t}, \qquad \text{or} \qquad \frac{J_{1t}}{\sigma_1} = \frac{J_{2t}}{\sigma_2}. \qquad (10.27)$$
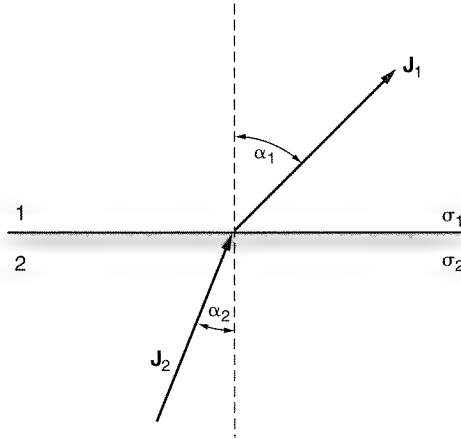
**Figure 10.6** Boundary surface between two conducting media
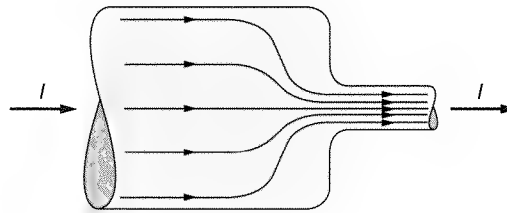


**Figure 10.7** The normal component of vector **J** in a conductor adjacent to an insulator is zero.

**Example 10.4—Boundary conditions between a conductor and an insulator.** If one of the two media in Fig. 10.6, say, medium 1, is an insulator (e.g., air), what are the expressions for the boundary conditions?

We know that there can be no current in the insulator. Hence from Eq. (10.26) we conclude that the normal component of the current density vector in a conductor adjacent to the insulator must be zero. Tangential components of the vector **E** are the same in the two media, but of course $J_{1t}$ does not exist. Lines of vector **J** in such a case are sketched in Fig. 10.7.

*Questions and problems:*    Q10.13

## 10.7   Grounding Electrodes and an Image Method for Currents

Consider an electrode (a conducting body) of arbitrary shape buried in a poorly conducting medium (for example, soil) near the flat boundary surface between the poor conductor and some insulator (for example, air). Suppose that a current of intensity $I$ is flowing from the electrode into the conducting medium, and is supplied from a distant current source through a thin insulated wire, as shown in Fig. 10.8. According to the boundary conditions, the current flow lines are tangential to the surface.
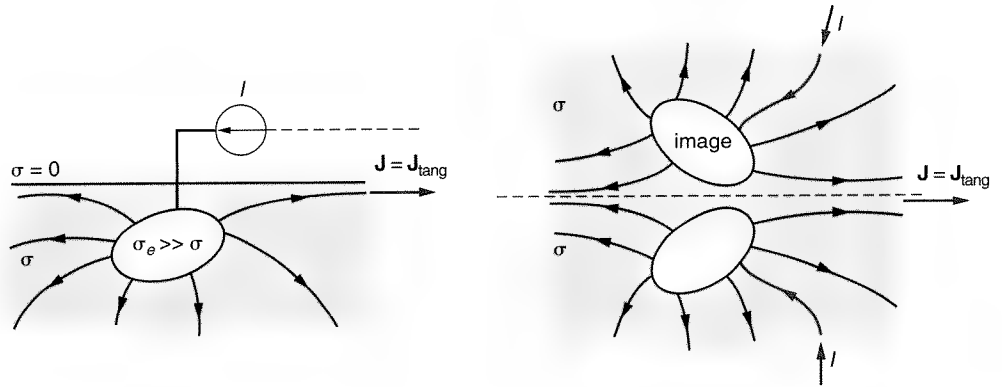
Figure 10.8 Image method for dc currents flowing out of an electrode into a poor conductor near the interface with an insulator

Imagine that the entire space is filled with the poor conductor. If an image electrode *with the current of the same intensity flowing out of it* is placed in the upper half of the space, as shown in the figure, the resulting current flow in both half spaces will be tangential to the plane of symmetry (the former boundary surface). Therefore, the current distribution in the lower half space is the same as in the actual case. Consequently, the influence of the boundary surface may be replaced by the image electrode, provided the current through the image electrode has the same intensity *and direction* as that through the real electrode. This method is very useful for analyzing current flow from electrodes buried under the surface of the earth. Such electrodes are often used for grounding purposes.

**Example 10.5—Hemispherical grounding electrode.** Suppose that a hemispherical electrode of very high conductivity is buried in poorly conducting soil of conductivity $\sigma$, as shown in Fig. 10.9. We are interested in
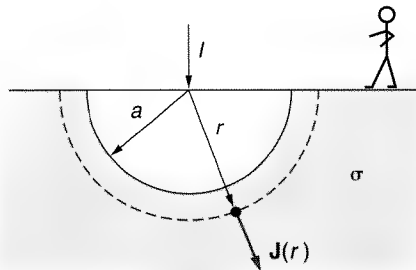


Figure 10.9 A hemispherical grounding electrode

- The resistance of such a grounding system
- The intensity of the electric field at all points on the earth's surface if a current $I$ flows from the electrode into the earth
- What happens if a person in noninsulating shoes approaches this grounding electrode

We use the image method. Let the current $I$ flow out of the hemispherical electrode. If all space is filled with earth, the image is another hemispherical electrode, so we get a spherical electrode with current $2I$ in a homogeneous conducting medium. Due to symmetry, the current from the spherical electrodes is radial. Consequently, the current from the original hemispherical electrode is also radial. The current density at a distance $r$ from the center of the hemisphere is therefore

$$J = \frac{2I}{4\pi r^2} = \frac{I}{2\pi r^2}.$$

The magnitude of the electric field is

$$E = \frac{J}{\sigma} = \frac{I}{2\pi \sigma r^2},$$

so that the potential of the electrode with respect to a reference point at infinity is obtained as

$$V_a = \int_a^\infty E \, dr = \int_a^\infty \frac{I}{2\pi \sigma r^2} dr = \frac{I}{2\pi \sigma a}.$$

Is it possible to define the resistance of this grounding system? We need two terminals for a resistor, and the hemispherical electrode has (seemingly) just one. However, the current must be collected somewhere at a distant point by a generator and returned to the electrode through a wire. The distant point is the other "resistor" contact. Usually this other contact is a large grounding system of a power plant, with a large contact area with the ground, so that the principal contribution to the resistance of this resistor comes from the hemispherical electrode. We can therefore define the resistance of the hemispherical grounding electrode as

$$R = \frac{V_a}{I} = \frac{1}{2\pi \sigma a}.$$

The potential at any point on the surface of the earth due to the current flow is

$$V_r = \int_r^\infty E(r) dr = \frac{I}{2\pi \sigma r},$$

and the potential difference between two points at a distance $\Delta r = d$ apart is given by

$$\Delta V = \frac{I}{2\pi \sigma r} - \frac{I}{2\pi \sigma (r + d)}.$$

Suppose a person approaches the grounding electrode when a large current of $I = 1000$ A is flowing through it. The potential difference between his feet can be very large and even fatal. For example, if $\sigma = 10^{-2}$ S/m, $r = 1$ m, and the person's step is $d = 0.75$ m long, the potential difference between the two feet will be 6820 volts.

Real grounding electrodes are usually in the shape of a plate, a rod, or a thin metal mesh and are buried in the ground. The variation in the conductivity of the soil has a large effect on the behavior of the grounding electrode. For that reason the conductivity around it is sometimes purposely increased by, for example, adding salt to the soil. The example of a hemispherical electrode is useful because it gives us an idea of the order of magnitude, but it is certainly not precise.

*Questions and problems:*   Q10.14 to Q10.17, P10.17 to P10.19

## 10.8   Chapter Summary

1. The basic quantities that describe electric current are the current density vector, **J**, describing current flow *at any point*, and the current intensity, *I*, *describing current flow through a surface.*

2. The current density vector **J** is most frequently a linear function of the local electric field strength, $\mathbf{J} = \sigma \mathbf{E}$ (point form of Ohm's law), also expressed as $\mathbf{E} = \rho \mathbf{J}$, where $\sigma$ is the conductor conductivity, and $\rho$ its resistivity.

3. The volume power density of transformation of electric energy into heat in conductors is described by the point form of Joule's law, $p_J = \mathbf{J} \cdot \mathbf{E}$.

4. The current continuity equation is a mathematical expression of the experimental law of conservation of electric charges. Its form for time-invariant currents is just the Kirchhoff's current law of circuit theory.

5. A resistor is an element, made of a resistive material, with two terminals, each equipotential. Ohm's and Joule's laws for resistors known from circuit theory are derived from field theory.

6. It is not possible to maintain a current field without devices known as electric generators, which convert some other form of energy into electric energy, i.e., energy of separated electric charges. Electric generators are characterized by their electromotive force and internal resistance.

7. Based on boundary conditions for current fields, an image method can be formulated for these fields, similar to that in electrostatics. However, in this case the boundary conditions are satisfied by a "positive" image. The image method can be used, for example, to determine the field and resistance of grounding electrodes.

## QUESTIONS

**Q10.1.** What do you think is the main difference between the motion of a fluid and the motion of charges constituting an electric current in conductors?

**Q10.2.** Describe in your own words the mechanism of transformation of electric energy into heat in current-carrying conductors.

**Q10.3.** Is Eq. (10.5) valid also for a closed surface, or must the surface be open? Explain.

**Q10.4.** A closed surface $S$ is situated in the field of time-invariant currents. What is the charge that passes through $S$ during a time interval $dt$?

**Q10.5.** Is a current intensity on the order of 1 A frequent in engineering applications? Are current densities on the order of $1\,\text{mA/m}^2$ or $1\,\text{kA/m}^2$ frequent in engineering applications? Explain.

**Q10.6.** What is the difference between linear and nonlinear resistors? Can you think of an example of a nonlinear resistor?

**Q10.7.** A wire of length $l$, cross-sectional area $S$, and resistivity $\rho$ is made to meander very densely. The lengths of the successive parts of the meander are on the order of the wire radius. Is it possible to evaluate the resistance of such a wire accurately using Eq. (10.19)? Explain.

**Q10.8.** Explain in your own words the statement in Eq. (10.20).

**Q10.9.** Assume that you made a resistor in the form of an uninsulated metal container (one resistor contact) with a conducting liquid (e.g., tap water with a small amount of salt), and a thin wire dipped into the liquid (the other resistor contact). If you change the level of water, but keep the length of the wire in the liquid constant, will this produce a substantial variation of the resistor resistance? If you change the length of the wire in the liquid, will this produce a substantial variation of the resistor resistance? Explain.

**Q10.10.** What is meant by "nonelectric forces" acting on electric charges inside electric generators?

**Q10.11.** List a few types of electric generators, and explain the nonelectric (impressed) forces acting in them.

**Q10.12.** Does the impressed electric field strength describe an *electric* field? What is the unit of the impressed electric field strength?

**Q10.13.** Prove that on a boundary surface in a time-invariant current field, $J_{1\text{norm}} = J_{2\text{norm}}$.

**Q10.14.** Where do you think the charges producing the electric field in the ground in Fig. 10.9 are located? (These charges cause the current flow in the ground.)

**Q10.15.** Explain in your own words what is meant by the grounding resistance, and what this resistor is physically.

**Q10.16.** Is it possible to define the grounding resistance if the generator is not grounded? Explain.

**Q10.17.** Assume that a large current is flowing through the grounding electrode. Propose at least three different ways to approach the electrode with a minimum danger of electric shock.

# PROBLEMS

**P10.1.** Prove that the current in any homogeneous cylindrical conductor is distributed uniformly over the conductor cross section.

**P10.2.** Uniformly distributed charged particles are placed in a liquid dielectric. The number of particles per unit volume is $N = 10^9\,\text{m}^{-3}$, and each is charged with $Q = 10^{16}\,\text{C}$. Calculate the current density and the current magnitude obtained when such a liquid moves with a velocity of $v = 1.2\,\text{m/s}$ through a pipe of cross section area $S = 1\,\text{cm}^2$. Is this current produced by an electric field?

**P10.3.** A conductive wire has the shape of a hollow cylinder with inner radius $a$ and outer radius $b$. A current $I$ flows through the wire. Plot the current density as a function of radius, $J(r)$. If the conductivity of the wire is $\sigma$, what is the resistance of the wire per unit length?

**P10.4.** A conductor of radius $a$ is connected to one with radius $b$. If a current $I$ is flowing through the conductor, find the ratio of the current densities and of the densities of Joule's losses in both parts of the conductor if the conductivity for both parts is $\sigma$.

**P10.5.** The homogeneous dielectric inside a coaxial cable is not perfect. Therefore, there is some current, $I = 50\,\mu A$, flowing through the dielectric from the inner toward the outer conductor. Plot the current density inside the cable dielectric, if the inner conductor radius is $a = 1\,mm$, the outer radius $b = 7\,mm$, and the cable length $l = 10\,m$.

**P10.6.** Find the expression for the current through the rectangular surface $S$ in Fig. P10.6 as a function of the surface width $x$.
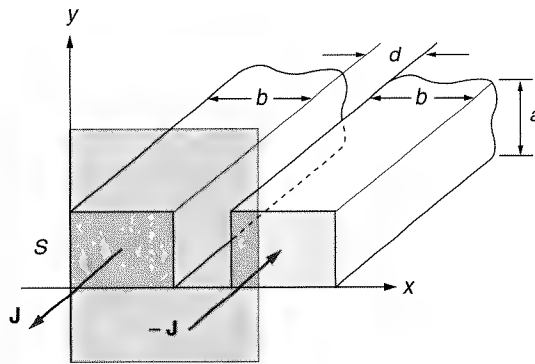


Figure P10.6 Calculating current intensity

**P10.7.** Find the expression for the current intensity through a circular surface $S$ shown in Fig. P10.7 for $0 < r < \infty$.
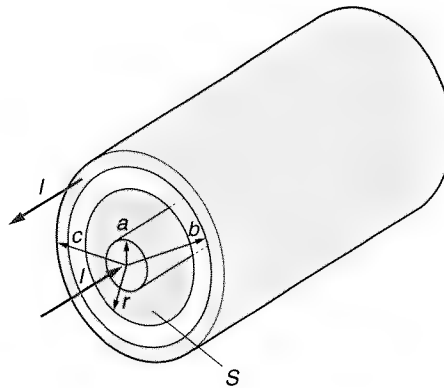


Figure P10.7 A coaxial cable

**P10.8.** The resistivity of a wire segment of length $l$ and cross-sectional area $S$ varies along its length as $\rho(x) = \rho_0(1 + x/l)$. Determine the wire segment resistance.

**P10.9.** Find the resistance between points 2 and 2′ of the resistor shown in Fig. P10.9.
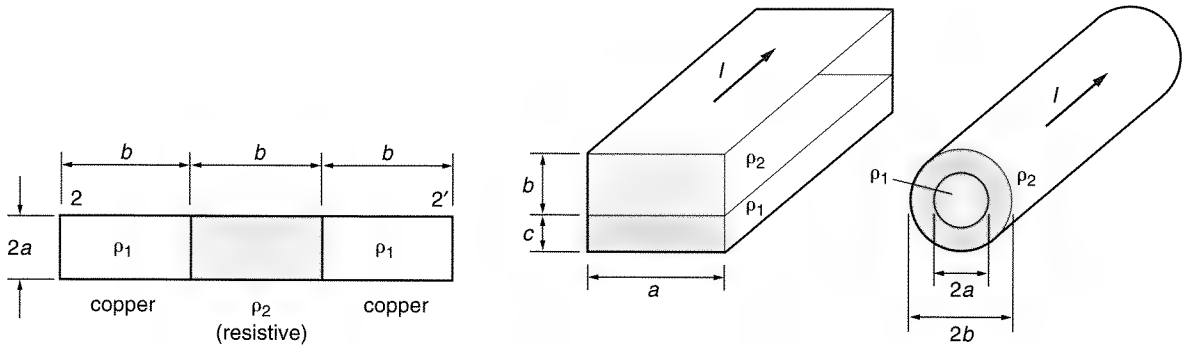


Figure P10.9 An idealized resistor CROSSECTION    Figure P10.10 Two inhomogeneous conductors
OF RADIUS $a$.

**P10.10.** Show that the electric field is uniform in the case of both inhomogeneous conductors in Fig. P10.10. Find the resistance per unit length of these conductors, the ratio of currents in the two layers, and the ratio of Joule's losses in the two layers.

**P10.11.** The dielectric in a coaxial cable with inner radius $a$ and outer radius $b$ has a very large, but finite, resistivity $\rho$. Find the conductance per unit length between the cable conductors. Specifically, find the conductance between the conductors of a cable $L = 1$ km long with $a = 1$ cm, $b = 3$ cm, and $\rho = 10^{11} \, \Omega \cdot$ m.

**P10.12.** A lead battery is shown schematically in Fig. P10.12. The total surface area of the lead plates is $S = 3.2 \, \mathrm{dm}^2$, and the distance between the plates is $d = 5$ mm. Find the approximate internal resistance of the battery, if the resistivity of the electrolyte is $\rho = 0.016 \, \Omega \cdot$ m.
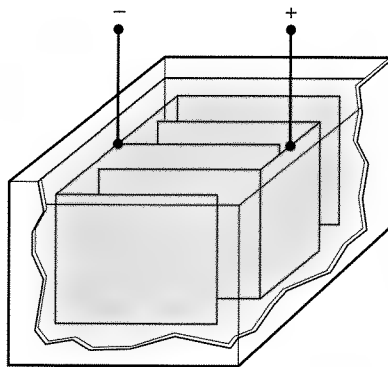


Figure P10.12 A lead battery

**P10.13.** Calculate approximately the resistance between cross sections 1 and 2 of the nonuniform strip conductor sketched in Fig. P10.13. The resistivity of the conductor is $\rho$. Why can the resistance be calculated only approximately?
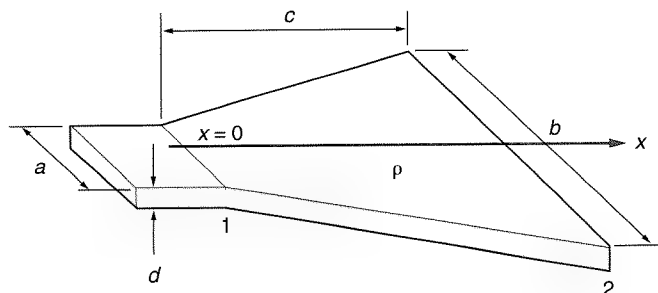


Figure P10.13 A nonuniform strip conductor

**P10.14.** Calculate approximately the resistance between cross sections 1 and 2 of the conical part of the conductor sketched in Fig. P10.14. The resistivity of the conductor is $\rho$.
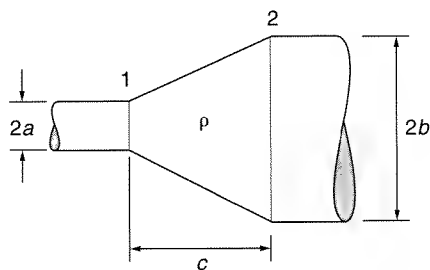


Figure P10.14 A conical conductor
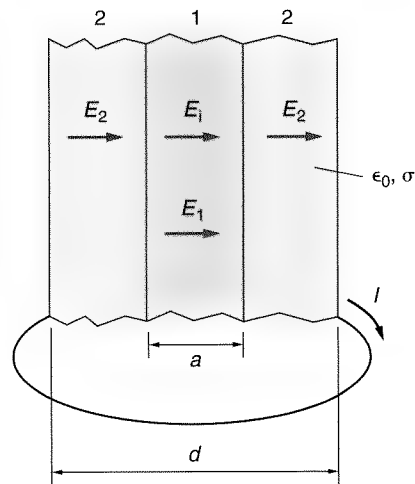
Figure P10.15 A resistor with an impressed field

**P10.15.** In the darker shaded region of the very large conducting slab of conductivity $\sigma$ and permittivity $\epsilon_0$ shown in Fig. P10.15, a uniform impressed field $E_i$ acts as indicated. The end surfaces of the slab are coated with a conductor of conductivity much greater than $\sigma$, and connected by a wire of negligible resistance. Determine the current density, the electric field intensity, and the charge density at all points of the system. Ignore the fringing effect.

**P10.16.** Repeat problem P10.15 assuming that the permittivity of the slab is $\epsilon$ (different from $\epsilon_0$). Note that in that case polarization charges are also present.

**P10.17.** Determine the resistance of a hemispherical grounding electrode of radius $a$ if the ground is not homogeneous, but has a conductivity $\sigma_1$ for $a < r < b$, and $\sigma_2$ for $r > b$, where $b > a$, and $r$ is the distance from the grounding electrode center.

**P10.18.** Determine the resistance between two hemispherical grounding electrodes of radii $R_1$ and $R_2$, which are a distance $d$ ($d \gg R_1, R_2$) apart. The ground conductivity is $\sigma$.

**P10.19.** A grounding sphere of radius $a$ is buried at a depth $d$ ($d \gg a$) below the surface of the ground of conductivity $\sigma$ (Fig. P10.19). Determine the points at the surface of the ground at which the electric field intensity is the largest. Determine the electric field intensity at these points if the intensity of the current through the grounding sphere is $I$.
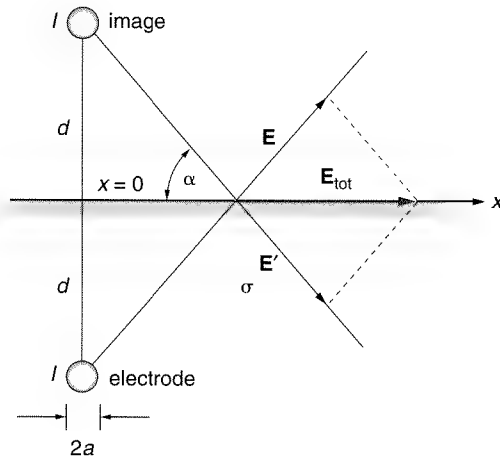
Figure P10.19 A deeply buried spherical electrode

# 11

# Some Applications
# of Electrostatics

## 11.1  Introduction

At this point, we will briefly study some common instances and applications of electrostatic fields. The earth's strongest electric field is the atmospheric field between charged clouds, and between charged clouds and the ground. We will briefly describe fields, voltages, and currents for the extreme case of thunderstorms. Commercial applications of electrostatics include pollution-control filters, xerography, printing, electrostatic separation, coating, and some medical and biological applications. We will also look at some applications where understanding of dc current density is needed, such as probes for characterizing semiconductor materials. We will review only the basic principles of these processes and include some examples of engineering design parameters that directly apply the knowledge gained in the previous chapters. Even at that level, the examples in this chapter will show clearly how powerful the knowledge we have gained is for understanding existing devices and for discovering and designing new ones.

159

## 11.2   Atmospheric Electricity and Storms

Thunderstorms are the most obvious manifestations of electrical phenomena on our planet. A typical storm cloud carries about 10 to 20 coulombs of each type of charge, at an average height of 5 km above the earth's surface. Where does the electric energy of a cloud come from? The main source of energy on our planet is the sun: huge water masses on earth are heated by the sun's energy. The evaporated water, which contains energy originated from the sun, forms clouds. A small portion of the energy of a cloud is turned into electric energy. So thunderstorms are huge (but quite inefficient) thermoelectric generators. Some of the suspected mechanisms of cloud electrification are the breakup of raindrops, freezing, frosting, and friction between drops or crystals. Movements inside a cloud eventually make the cloud into an electric dipole, with negative charges in the lower part and positive charges in the upper part. The lower part induces positive charges on the earth below the cloud (Fig. 11.1a). The induced charge density is greatest on tall, sharp objects.
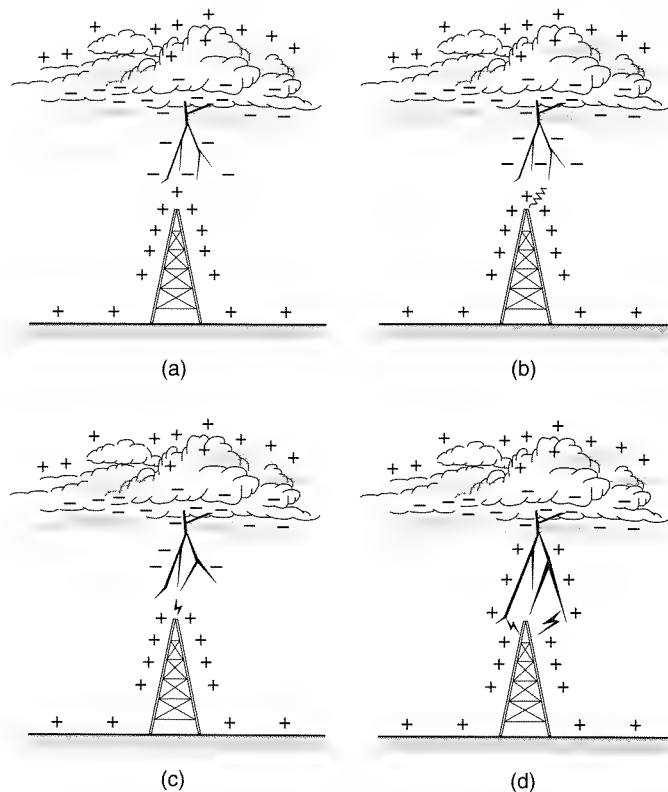


Figure 11.1 Formation of a lightning bolt. (a, b) The stepped-leader down-flowing stroke defines the path for the visible return stroke (c, d).

The beginning of a cloud-to-ground lightning bolt is an invisible discharge, called the *stepped leader*. The stepped-leader air breakdown is initiated at the bottom of the cloud. It moves in discrete steps, each about 50 m long and lasting for about 1 $\mu$s. Because of this discharge, electrons are released from the lower part of the cloud. These electrons are attracted by the induced positive charges on the ground, and they move downward (Fig. 11.1b). As the negatively charged stepped leader approaches the ground, it induces even more positive charges, especially on protruding objects. When the leader is about 100 m above the ground, a spark moves up from the ground to meet it. Once a conductive path is established, huge numbers of electrons flow from the cloud to the ground. To balance the charge flow, positive charges move up toward the cloud, trying to neutralize the huge negative charge at the bottom of the cloud. This is the discharge that we see and is called the *return stroke*. The return stroke lasts only about 100 $\mu$s, so it looks as if the entire channel lights up at the same time.

The currents in the return stroke are typically on the order of 10 kA but can be as large as 200 kA. The temperatures associated with the Joule's heat of such a current are very high—temperatures in the return stroke reach 30,000 K. The air does not have time to expand in volume, so its pressure rises to several million pascals, causing a sound shock—thunder.

At any time, there are about 1800 thunderstorms across the earth, and about 100 lightning flashes per second. The National Center for Health Statistics estimates that the death toll of lightning, about 150 people every year in the United States, is bigger than that of hurricanes and tornadoes. Lightning also causes considerable damage. Lightning rods (connected to grounding electrodes) protect exposed structures from damage by routing the strokes to the ground through the rod rather than through the structure. Benjamin Franklin suggested the use of lightning rods, and they were in place in the United States and France as early as the middle of the 18th century. It is estimated that for rural buildings containing straw and protected by a rod, the danger of fire caused by lightning is reduced by a factor of 50.

*Questions and problems:*    Q11.1 to Q11.3, P11.1 and P11.2

## 11.3   Electric Current in a Vacuum and in Gases

In lightning, current flows from the clouds to the earth through atmosphere, which is a gas. There are two major differences between electric currents in solid and liquid conductors and those in a vacuum and in rarefied (low-pressure) gases. First, the electric charges in a vacuum or in gases are most often moved by the electric field of some stationary charges only; there is usually no impressed electric field. Second, in a vacuum and in gases there is no relation similar to the point form of Ohm's law. This is obvious in the case of a vacuum: charges do not collide with atoms but move under the influence of the electric forces and forces of inertia only. Consequently they do not follow the lines of vector **E** except if the field lines are straight.

In the case of gases, particularly if rarefied, an accelerated ion has a relatively long path between collisions, so it can acquire a considerable kinetic energy. As a consequence, in rarefied gases Ohm's law is not valid. Various other effects can oc-

cur, however, due to the possibility of a chain production of new pairs of ions by collisions of high-velocity ions with neutral molecules. Electrostatic discharge due to high voltages is one such effect that is of practical interest and will be described briefly in this chapter.

**Example 11.1—Motion of electric charges in the electric field.** Let a charge $Q$ of mass $m$ move in an electrostatic field in a vacuum. The equation of motion of the charge has the form

$$m\frac{d^2\mathbf{r}(t)}{dt^2} = Q\mathbf{E}(\mathbf{r}), \tag{11.1}$$

where $\mathbf{r}(t)$ is the position vector (variable in time) of the charge, and $\mathbf{E}$ is the electric field strength, a function of coordinates (that is, of $\mathbf{r}$).

Let us now look specifically at the motion of a charged particle in a uniform electric field. Assume that a charge $Q$ ($Q < 0$) leaves with a negligibly small velocity the negative plate of a charged parallel-plate capacitor in which the electric field strength is $E$. Let the plates be a distance $d$ apart. We wish to determine the position and velocity of the charge as a function of time, if the charge left the plate at $t = 0$.

Let the $x$ axis be perpendicular to the plates, with $x = 0$ at the negative plate (Fig. 11.2). The charge will move parallel to the $x$ axis. According to Eq. (11.1), the equation of motion is

$$m\frac{d^2x}{dt^2} = QE,$$

that is, the charge moves with constant acceleration $QE/m$. The charge velocity as a function of time is given by

$$v = \frac{dx}{dt} = \frac{QE}{m}t,$$
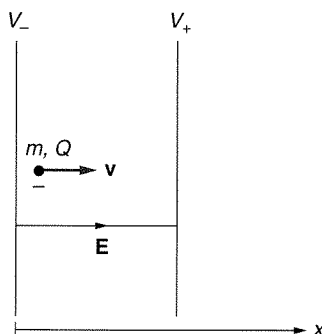


Figure 11.2 Motion of a charged particle in a uniform electric field between two parallel charged plates

since the initial velocity is zero. From this equation, the position of the particle is given by

$$x = \frac{QE}{2m} t^2.$$

Assume that the voltage between the plates is $V = 1000$ V, that the charge is an electron ($e \simeq -1.6 \cdot 10^{-19}$ C, $m \simeq 9.1 \cdot 10^{-31}$ kg), and that the distance between the plates is $d = 10$ cm. The electric field strength between the plates is then $E = V/d = 10,000$ V/m.

The charge will reach the positive plate after a time obtained from the last equation, in which we set $x = d$. This yields $t = 1.07 \cdot 10^{-8}$ s. The velocity of the electron before impact with the positive electrode is now obtained if we insert this particular time into the expression for the velocity. The result is $v = 18.8 \cdot 10^6$ m/s, a very large velocity (close to those velocities for which relativistic corrections in particle mass need to be made). In conclusion, electrons can be accelerated to remarkably high velocities with voltages that are not difficult to produce.

The preceding equations were derived for a uniform electric field. For arbitrary $\mathbf{E}(\mathbf{r})$ no analytical solution of Eq. (11.1) for the position vector $\mathbf{r}$ in time is known, but it can always be found approximately by numerical methods. Often the exact trajectory of the particle is not of interest. Instead only the magnitude of the velocity of the particle is required, and it can easily be calculated.

Let the charged particle (of charge $Q$ and mass $m$) leave point 1, which is at potential $V_1$, with a velocity of magnitude $v_1$. We wish to determine the magnitude $v_2$ of its velocity when it reaches point 2, which is at a potential $V_2$. In moving the particle from point 1 to point 2, the electric forces perform work $Q(V_1 - V_2)$, and due to energy conservation, we know that the particle kinetic energy must have increased by precisely that amount. Thus

$$\frac{mv_2^2}{2} - \frac{mv_1^2}{2} = Q \int_1^2 \mathbf{E} \cdot d\mathbf{l} = Q(V_1 - V_2). \tag{11.2}$$

The magnitude of the velocity at point 2 is

$$v_2 = \sqrt{v_1^2 + \frac{2Q(V_1 - V_2)}{m}}. \tag{11.3}$$

In the particular, but common, case when $v_1 = 0$, this becomes

$$v_2 = \sqrt{\frac{2Q(V_1 - V_2)}{m}}. \tag{11.4}$$

(*Velocity of a charge accelerated from zero velocity by potential difference $V_1 - V_2$*)

**Example 11.2—Velocity of an electron accelerated by 1 kV.** As a numerical example, let us determine the velocity of an electron accelerated from zero velocity by a 1000-V voltage. Using Eq. (11.4) we get $v = 18.8 \cdot 10^6$ m/s, as in Example 11.1 for the special case of a uniform field. Note that this result is valid for *any* electric field (not necessarily uniform), in which the electron covers a voltage of 1000 V.

*Questions and problems:*   Q11.4 to Q11.9, P11.3 to P11.7

## 11.4   Corona and Spark Discharge

In gases, and particularly in air at normal atmospheric pressure, specific steady discharging currents may occur in certain circumstances. A necessary condition for this process is a region of electric field with intensity greater than the dielectric strength of air (about 30 kV/cm). In this region the air becomes ionized, i.e., conducting, which is equivalent to an enlargement of the electrode. If the electric field intensity outside this enlarged "electrode" is less than the dielectric strength of air, the process stops. The cloud of charges around the electrode stays permanently, and it forms a source of ions that are propelled toward the electrodes of the opposite sign. As a result, there are steady discharging currents between the electrodes. The ionized cloud is known as a *corona*.

In some instances, however, the electric field strength may not be decreased when a corona is formed around an electrode. The process then does not stop, but instead spans the whole region between the electrodes. Violent discharge of the electrodes occurs, known as *spark discharge*. The spark discharge is not, of course, a time-invariant phenomenon.

Normally corona is not desirable, because it results in losses of charge on charged conductors. In some instances, however, it is of great use. For example, discharging of an aircraft that is charged during flight by friction is performed by encouraging corona discharges at several positions on the aircraft. We will see in the next sections that corona discharge is done on purpose when small neutral metal or dielectric particles need to be charged, for example in electrostatic painting of cars.

Spark discharge is also usually undesirable. For example, in manufacturing plants or coal mines where explosive gases may exist, electrostatically charged bodies may discharge through sparks, which in turn may have enough energy to initiate a large-scale explosion. Spark discharge is a relatively frequent cause of explosions involving loss of human lives and property.

*Questions and problems:*    Q11.10 to Q11.12

## 11.5   Electrostatic Pollution-Control Filters

Electrostatic filters are used in environmental control for removing fine particles from exhaust gases. In the filtering (or precipitation) process, the particles are charged, separated from the rest of the gas by a strong electric field, and finally attracted to a pollutant-collecting electrode. In the United States there are a few thousand large industrial electrostatic filters, and a large number of small units used for indoor air cleaning. Electric power generation plants in the United States generate about $2 \times 10^7$ tons of coal fly ash every year. This ash accounts for most of the use of electrostatic filters, although steel and cement production, paper processing, sulfuric acid manufacturing, petroleum refining, and phosphate and other chemical processing also use electrostatic filters.

A simplified diagram of a filter (this type is referred to as the "tubular" precipitator) is shown in Fig. 11.3. The polluted gas flow enters the bottom of the cylinder and the small ash particles are charged by ionized air around a high-potential
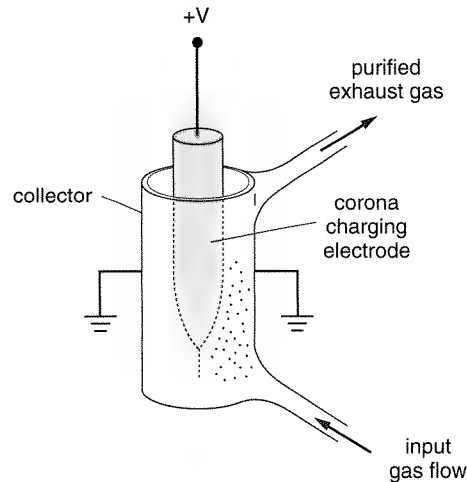
Figure 11.3 Simplified diagram of an electrostatic tubular filter

electrode (to learn more about this mechanism of gas ionization, the reader is encouraged to refer to, e.g., A.D. Moore, ed., *Electrostatics and its applications*, John Wiley, 1973). The electric field between the high-potential electrode and the coaxial grounded cylinder causes the charged particles to be attracted to the cylinder, which acts as the collecting electrode. The purified gas flows through the clean-gas exhaust at the top of the cylinder.

To understand the engineering design problems in electrostatic filters, let us consider first an uncharged idealized spherical conductive particle in a uniform electric field (Fig. 11.4a). Due to the presence of the external field, induced charges are created on the surface of the sphere. Using knowledge we gained in Chapter 6, it can be shown that the resulting electric field at points $A$ and $B$ on the sphere is three times that of the external field (see Example 11.3).

This means that the uniform external field has been changed by the presence of the uncharged particle and is now nonuniform, as shown in Fig. 11.4a. (The field inside the particle is not shown, because it is different for conductive or dielectric particles, as will be discussed shortly.) We said that in an electrostatic filter these particles encounter ionized air when they enter the electric field. This means that they are not strictly in a vacuum, since there are occasional charged ions in the space around them. These ions will be attracted to the particle if they are found in a certain region close to it, as shown in Fig. 11.4a. In this case, the particle becomes charged, and the excess charge distributes itself to satisfy the boundary conditions. This has an effect on the surrounding field, which becomes more nonuniform than in the previous case, as shown in Fig. 11.4b. The result is that it will be more difficult for the next ion of the same sign to be attracted to the particle, because the size of the region where it needs to find itself in order to be attracted is reduced.

The particles will be collected faster if they carry more charge, because then the electric force is larger. However, we have just explained that particles cannot be
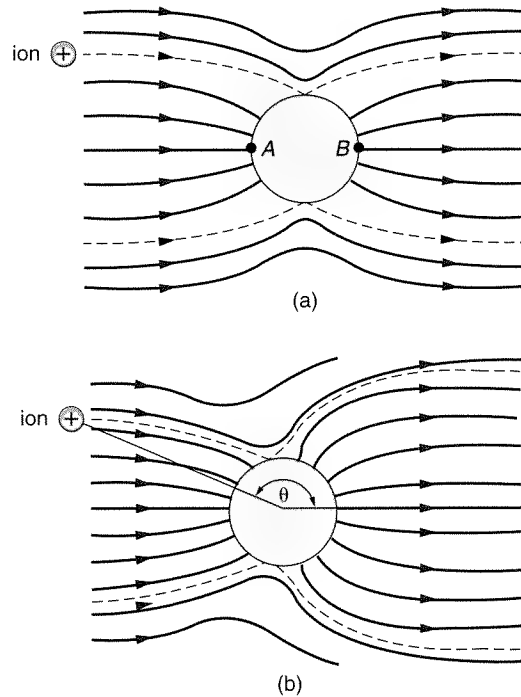
Figure 11.4 (a) An uncharged spherical dielectric pollutant particle and (b) a partly charged particle in a uniform electric field. A positive ion will be attracted to the particle in case (a) and repelled by it in case (b).

charged beyond a certain amount. From this relatively simple example, we can see that there is a limitation on how fast pollutant particles can be charged and collected inside an electrostatic filter. In order to speed up the process of collection, and therefore filter out as many pollutant particles as possible in a limited region of exhaust-gas flow inside the filter, different designs than the one shown in Fig. 11.4 have been in use. More details can be found in H.J. White, *Industrial electrostatic precipitation*, Addison-Wesley, 1963.

**Example 11.3—Dielectric and conductive spherical particles in a uniform electric field.** A dust particle can be approximated by a dielectric or conductive sphere in a (locally) uniform field. Consider first a uniformly polarized sphere of radius $a$. Let the polarization vector in the sphere be $\mathbf{P}$ (Fig. 11.5a). We assume for the moment that the sphere is situated in a vacuum, and that its polarization is the only source of the field.

The field outside and inside the sphere is the same as the field resulting only from the surface polarization charges. (Since $\mathbf{P}$ is constant, $\rho_p = \mathrm{div}\mathbf{P} = 0$.) The density of these surface charges is given by Eq. (7.16), which in the case considered becomes

$$\sigma_p = \mathbf{P} \cdot \mathbf{n} = P\cos\theta. \tag{11.5}$$
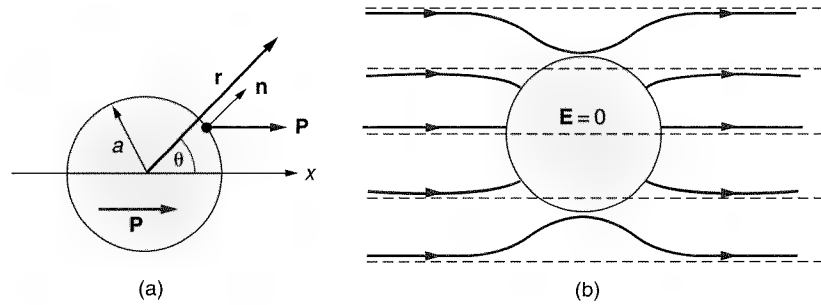
Figure 11.5 (a) A uniformly polarized dielectric sphere and (b) an uncharged conductive sphere in a uniform electric field

The field and potential of this charge distribution are not easy to find directly. However, we can use the following reasoning: the uniformly polarized dielectric sphere is equivalent to two spherical charged clouds of uniform volume densities $\rho$ and $-\rho$ whose centers are displaced by a small distance $d$. We already know that the field and potential outside such a system are equal to those of a dipole. This gives us the field and potential outside the dielectric sphere. The potential of a dipole is given by

$$V(r, \theta) = \frac{p \cos \theta}{4\pi \epsilon_0 r^2},$$

so all we need to find is the equivalent dipole moment $p$ for the two spherical clouds. Recall the definition of the polarization vector:

$$\mathbf{P} = \frac{\sum d\mathbf{p}_{\text{in } dv}}{dv}.$$

Since all the $\mathbf{p}$ moments in $dv$ are parallel, and $dp = (\rho\, dv)d$, we get

$$P = \rho d = \frac{Q}{4a^3\pi/3}d = \frac{p}{4a^3\pi/3}.$$

So the dipole moment of the two displaced charged spheres is

$$p = \frac{4}{3}a^3\pi P, \tag{11.6}$$

and thus we know the potential (and hence also the field) outside the uniformly polarized sphere.

We wish now to determine the potential (and the field) inside the sphere. Consider the potential

$$V(r, \theta) = \frac{pr \cos \theta}{4\pi \epsilon_0 a^3} = \frac{px}{4\pi \epsilon_0 a^3} \qquad (r \leq a). \tag{11.7}$$

It is a simple matter to prove that this potential satisfies Laplace's equation (so it is a physically possible potential). For $r = a$, it becomes identical with the potential outside the sphere. Since the potential is continuous across the sphere surface, we conclude that boundary conditions

are satisfied, and that the expression in Eq. (11.7) represents the potential inside the uniformly polarized sphere.

The electric field inside the dielectric sphere has only an $x$ component, and is given by

$$E_x = -\frac{\partial V}{\partial x} = -\frac{p}{4\pi\epsilon_0 a^3} = -\frac{P}{3\epsilon_0}. \tag{11.8}$$

MINUS

Thus, *the electric field inside the uniformly polarized sphere is uniform*, and in the $x$ direction.

Consider now a dielectric sphere in a uniform *external* electric field $E_0\mathbf{u}_x$. This field, of course, tends to uniformly polarize all dielectric bodies in it. However, polarization charges for irregular bodies will produce an irregular secondary field, and generally the polarization of the bodies will not be uniform. Only if the shape of the body is such that a uniform polarization of the body results in a uniform secondary electric field inside it will the polarization of the body in the end also be uniform. We have just demonstrated that for a dielectric sphere this is precisely the case. Thus, inside a dielectric sphere in a uniform field the total field is uniform.

To determine the polarization of the sphere and then its secondary field using the preceding equations, note that $\mathbf{P} = (\epsilon - \epsilon_0)\mathbf{E}$. In this case, $\mathbf{E}$ is the total electric field inside the sphere, equal to the sum of the external field, $E_0\mathbf{u}_x$, and the field in Eq. (11.8). It is left to the reader as an exercise to determine the polarization of the sphere, and hence the total field inside and outside the sphere.

In some cases, pollutant particles are conductive rather than insulating. For example, a dust particle can be approximated by a conductive sphere, as shown in Fig. 11.5b. When such an uncharged metal sphere is placed in a uniform electric field $\mathbf{E}$, charges are induced on its surface to cancel the electric field inside the sphere. Since the external field is uniform, the charges have to distribute themselves to produce an opposite uniform field inside the sphere. From the previous discussion of a dielectric sphere in a uniform electric field, we know that this charge distribution has to be of the form in Eq. (11.5). The electric field strength in Eq. (11.8) is $E$. Consequently, the induced surface charge on the sphere is of density

$$\sigma(\theta) = 3\epsilon_0 E \cos\theta. \tag{11.9}$$

Since $D_n = \epsilon_0 E_n = \sigma$, the largest value of the field on the surface of the ball is for $\theta = 0$ and $\theta = \pi$ (points $A$ and $B$ in the figure):

$$E_A = E_B = 3E. \tag{11.10}$$

At points $A$ and $B$ on the surface of a conducting dust particle, the electric field is *three times stronger* than the original uniform field.

*Questions and problems:*   Q11.13 and Q11.14, P11.8

## 11.6   Electrostatic Imaging—Xerography

The modern photocopy machine was invented by the physicist and lawyer Chester Carlson in 1938. In his patent work he saw the need for an inexpensive and easy way to copy documents. It took him about 10 years to develop the copier, and in 1947 the Haloid Company—now Xerox Corporation—licensed the invention and began commercial production. The first copier was introduced to the market in 1950. Carlson called the process *xerography* from the Greek words *xeros* "dry," and *graphos*,

"writing." Xerography uses a photosensitive material called selenium. Selenium is normally a dielectric, but when illuminated it becomes conductive. Some other materials, such as zinc oxide and anthracene, also have this property.

The copying process essentially has five steps, shown in Fig. 11.6. In the first step, a selenium-coated plate is charged evenly by sliding it under positively charged wires. An image of the document is then exposed onto the plate by a camera lens. In places where the plate is illuminated (corresponding to the white areas of the document), the selenium becomes a conductor and the charge flows away to metal contacts on the side of the plate. In other places, corresponding to the dark (printed) areas
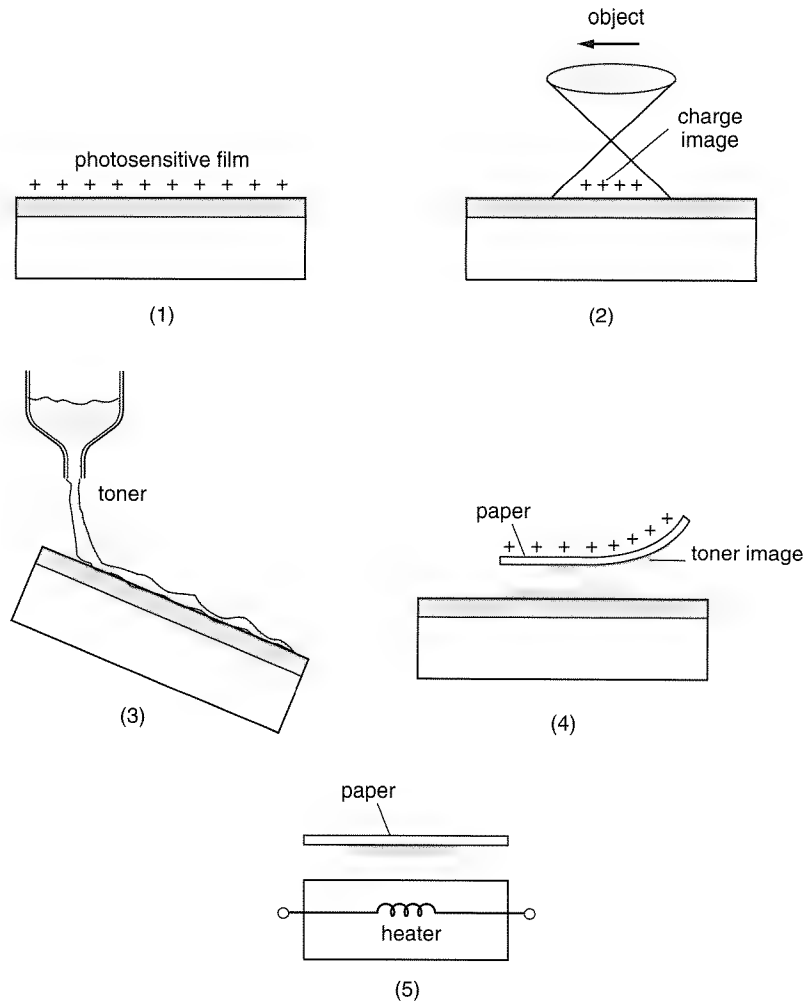


Figure 11.6 Five essential steps in the xerography process: (1) charging of the photoconductor plate; (2) charge image formation; (3) development with negatively charged toner; (4) transfer of toner image to paper; and (5) image fixing by heating

of the document, the charge remains. The plate now has an exact copy of the original in the form of a positive charge pattern. In the next step, some toner is charged negatively and the toner particles are attracted to the positively charged copy and stick to the plate. A sheet of blank paper is then placed over the plate and powder toner image. The paper is positively charged, so it attracts the toner particles onto itself. The paper is then placed on a fuser tray and the toner is baked to seal the image permanently. The entire process in the first copier took about 3 minutes. Modern copiers are faster and more sophisticated, but they operate according to the same principles.

What are some important electrostatic engineering design parameters in copiers? We said that the illuminated parts of the selenium plate become conductive and the charge flows away. But can the charge in the remaining charge image stay in place indefinitely, or do we need to time the next step in the copying process according to the temporal stability of the charge image? To answer this question, we need to know what the electric resistivity ($\rho$) and permittivity ($\epsilon$) of dark selenium (selenium with no illumination) are.

Once these two properties are known, we can reason in the following way. At the surface of the film, Gauss' law gives

$$\oint_{S_0} \mathbf{E} \cdot d\mathbf{S}_0 = \frac{1}{\epsilon} \int_S \sigma \mathbf{n} \cdot d\mathbf{S}, \tag{11.11}$$

where $\sigma$ is the surface charge density of the image on the film, $S_0$ is a thin, coinlike closed surface, with one base inside the film and the other over its very surface, and $S$ is the intersecting surface of $S_0$ and the film. The continuity equation can also be written for the current density $\mathbf{J} = \mathbf{E}/\rho$ flowing through the film due to its finite resistivity $\rho$:

$$\oint_{S_0} \mathbf{E} \cdot d\mathbf{S}_0 = -\rho \int_S \frac{\partial \sigma}{\partial t} \mathbf{n} \cdot d\mathbf{S}. \tag{11.12}$$

Eqs. (11.11) and (11.12) must be valid for any shape of the intersecting surface $S_0$. This is possible only if the expressions on their right sides are equal. In that case, the two equations can be combined to give a differential equation for the surface charge density, $\sigma$:

$$\frac{d\sigma}{dt} + \frac{1}{\epsilon\rho}\sigma = 0. \tag{11.13}$$

Assuming that at $t = 0$ the surface charge density is $\sigma_0$, the solution of this equation is $\sigma = \sigma_0 e^{-t/(\epsilon\rho)}$. The quantity $\epsilon\rho$ describes how quickly the charge image on the film diffuses. This quantity is called the *charge transfer time constant*, or *dielectric relaxation constant* of dark selenium. It tells us how long it takes for $\sigma_0/e \simeq 0.368\sigma_0$ of the charge in the image to flow away (e = 2.7182... is the base of natural logarithms). For selenium, the resistivity varies between $10^{11}\,\Omega \cdot m$ and $10^{14}\,\Omega \cdot m$, and the relative permittivity is about 6.1. The corresponding charge transfer time constants are 5.4 s to 5400 s (1.5 hours). This means that in the former case, after the charge image is formed by illumination, the toner image needs to be formed and transferred to paper in less than a few seconds in order to create a clear image.

Note that the charge transfer time constant $\epsilon\rho$ is in some sense similar to the $RC$ time constant of a resistor-capacitor circuit. We have seen before that the capacitance $C$ of a dielectric-filled capacitor is proportional to $\epsilon$, and the resistance $R$ is proportional to the resistivity $\rho$.

Another practical problem in copiers is developing, i.e., making a toner image from the charge image. The toner consists of small dielectric particles, about $10\,\mu$m in diameter, which are charged to about $Q = 0.5 \cdot 10^{-14}$ C. These particles are brought close to the photoconductive film, where a strong electric field, only a few times weaker than the air breakdown field, exists wherever there is a charge image. The electric force on a toner particle for a field strength three times smaller than the breakdown of air is then about

$$F = QE = 0.5 \cdot 10^{-14}\,\text{C} \cdot 10^6\,\text{V/m} = 0.5 \cdot 10^{-8}\,\text{N}. \tag{11.14}$$

If this force were the only one present, the toner particles would move exactly along the electric field vector lines. However, as we discussed in Example 11.3, if we assume that each particle is a tiny sphere, it becomes a dipole, and there is a force in addition to the force on the toner particles we just calculated, due to the inhomogeneous electric field of the charge image. After expressing the polarization vector $\mathbf{P}$ in Example 11.3 in terms of the permittivity of the dielectric, it can be shown that this additional force due to the field nonuniformity is about three orders of magnitude smaller than the force calculated in Eq. (11.14), and can therefore be neglected. The toner particles also have mass, and therefore a gravitational force is acting on them. This force is given by $F_g = 4\pi r^3 \rho g/3 \simeq 0.5 \cdot 10^{-11}$ N for a spherical particle $5\,\mu$m in radius and with mass density equal to that of water ($\rho = 10^3\,\text{kg/m}^3$). So the gravitational force can also be neglected, and our original conclusion that the toner particles follow the selenium-film electric field lines is a good approximation.

How is the toner brought to the charge image? Of several possible processes the following one is probably the simplest. A fine dust of carbon particles (on the order of $1\,\mu$m) is electrified and blown over the charge image, covering the charged parts of the image. If the paper surface has appropriate properties, the image is transferred onto the paper efficiently. Note that the small size of the dust particles enables a very high resolution image.

One of the engineering problems in this process is the fact that the electric field distribution around the edges of the image is nonuniform, as shown in Fig. 11.7a.
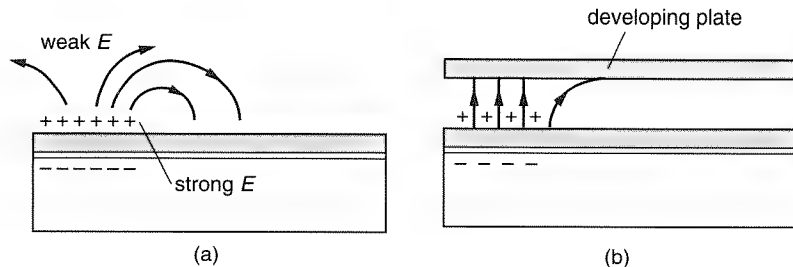


Figure 11.7 Field distribution around the edges of a charge image (a) without and (b) with the presence of a developing conductive plate

The field is the strongest around the edges, and toner particles are attracted only to the areas around the edges. In order to obtain toner coverage in the areas inside the image edges, a grounded conducting surface (called a developing electrode) is brought very near the charged film, but not touching it. The electric field strength between the photoconductive surface and the electrode is proportional to the surface charge density. The toner is introduced between the two surfaces, gets attracted to the surface charge on the film, and neutralizes this charge. The quantity of toner needed to effectively neutralize the charge at a certain spot on the film is proportional to the original surface charge density, which in turn is inversely proportional to the intensity of illumination during optical exposure. Thus the density of the developed image reproduces the continuous tones of the original optical image.

*Questions and problems:*   Q11.15 to Q11.18, P11.9

## 11.7   Industrial Electrostatic Separation

An important application of electrostatic fields is separation, used in industry for purification of food, purification of ores, sorting of reusable wastes, and sizing (sorting according to size and weight). Some specific examples of electrostatic separation in the ore and mineral industry are as follows: quartz from phosphates; diamonds from silica; gold and titanium from beach sand; limestone, molybdenite, and iron ore (hematite) from silicates; and zircon from beach sand. In the food industry, peanut beans, cocoa beans, walnuts, and nut meats are separated electrostatically from shells. For grains, electrostatics is used to separate rodent excrement from barley, soybean, and rice. In the electronics industry, copper wire is electrostatically separated from its insulation for recycling purposes. It is estimated that well over 10 million of tons of products a year are processed using electrostatic separation.

The first patent in this field was issued in 1880 for a ground cereal purification process. Thomas Edison had a patent in 1892 for electrostatically concentrating gold ore, and the first commercial process used in a plant in Wisconsin in 1908 was set up to electrostatically purify zinc and lead ores. Shortly after that, flotation processes were invented for separation. However, as these processes are not suitable for arid areas, and in some cases they require chemical reagents that present water pollution problems, electrostatic separation has regained popularity. In 1965, the world's largest electrical concentration plant was installed in the Wabush Mines in Canada. This plant is used to reduce the silica content of 6 million tons of iron ore per year.

Figure 11.8 shows two basic systems of electrostatic separation (there are several others not described here). In each case, the basic components of the system are a charging mechanism, an external electric field, a device to regulate the trajectory of nonelectric particles, and a feeding and collection system. In case (a) the particles are charged by contact electrification (the triboelectric effect), and in case (b) by ion bombardment (corona discharge). Both of these effects were mentioned earlier in this chapter. In terms of regulating the trajectory of the particles, in case (a) the gravitational force is used in addition to the electric force, and in case (b) the centrifugal force acting on the particles is adjusted by a rotating cylinder.
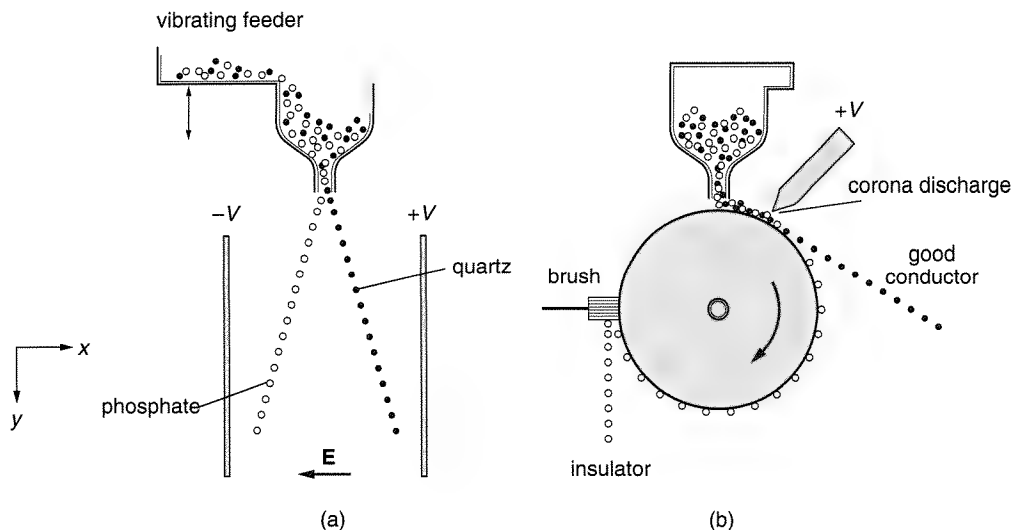
Figure 11.8 (a) Forming chute and (b) variable-speed rotating cylinder systems for electrostatic separation

In each case, physical separation of two types of particles is performed by adjusting the forces acting on the particles, as well as the time the forces act, so that different types of particles will have different trajectories. In Example 11.4, an approximate analysis of the separation process is examined for case (a) in Fig. 11.8. In case (b), used for separation of hematite from quartz or copper wire from its insulation, the particles are charged through ion bombardment. One type of particle is conductive, and the other dielectric. The particles are dropped on top of a grounded rotating cylinder. The conductor particles share their charge with the grounded rotor and are thrown from the rotor in a trajectory determined by centrifugal forces, gravity, and air resistance. The dielectric particles are held to the surface of the rotor. They either lose their charge slowly and then fall off the rotor, or are scraped off by a brush at the other end. The electric and centrifugal forces need to be roughly the same in order for the particles to stick to the rotor long enough for good separation.

**Example 11.4—Separation of quartz from phosphate rock using gravitational and electric forces.** An approximate analysis of the forming chute separation of quartz from phosphate rock can be done by ignoring the electric forces acting on particles due to neighboring charges. Obviously, this analysis is very rough and provides us only with an order-of-magnitude illustration.

The quartz and phosphate rock particles are washed, dried, heated to about 100°C, and finally vibrated so that the minerals make and break contact and are charged by friction with charges of opposite sign. For a uniform electric field acting along the $x$ direction in Fig. 11.8a, Newton's first law gives us an equality between the electric force on a single particle and its mass multiplied by its acceleration,

$$\mathbf{F}_e = Q\mathbf{E} = m\frac{d^2x}{dt^2}\mathbf{u}_x, \tag{11.15}$$

which can be integrated. Assuming zero initial velocity and displacement, the expression for the deflection of a charged particle due to the electric force is

$$x = \frac{1}{2}\frac{QE}{m}t^2.$$  (11.16)

For a 0.25-mm-diameter quartz particle, the ratio $Q/m \simeq 9 \times 10^{-6}$ C/kg, and a typical value for the electric field strength is $E = 4 \times 10^5$ V/m. Therefore, Eq. (11.16) gives $x = 1.8t^2$ m. The time required for a particle to fall a certain distance is obtained from the expression for the gravitational force

$$\mathbf{F}_g = m\mathbf{g} = -m\frac{d^2y}{dt^2}\mathbf{u}_y.$$  (11.17)

The vertical displacement of the particle due to the gravitational force is found by integration to be

$$y = -\frac{1}{2}gt^2.$$  (11.18)

For a falling distance of, say, 0.5 m, we can now calculate the time that the electric force has to deflect the particle in the horizontal direction as $t^2 \approx 0.1$ s$^2$, and the horizontal displacement as $d \approx 18$ cm. In the case of two particles that are oppositely charged, after falling 0.5 m they are separated by 36 cm, which is enough for good separation.

The engineering limitations of this process relate to the fact that the process can be used only for a certain range of particle sizes: if the particles become too large, roughly larger than 1 mm in diameter, the gravitational force becomes too large compared to the electric force. If the particles are too small, below roughly 50 $\mu$m, interparticle attractive electric forces cause the small quartz and phosphate particles to form clusters.

*Questions and problems:*   Q11.19, P11.10

## 11.8   Four-Point Probe for Resistivity Measurements

Four-point probes are used in every semiconductor lab. They can be used for determining the resistivity (conductivity) of a material, or the charge concentration if the other material properties are known. First consider just two probes (two points) that are touching the surface of a material of unknown conductivity, as in Fig. 11.9. The idea behind using probes is the same as measuring the resistance of a resistor, except that some of the quantities are distributed (i.e., described at every point, not only at the two resistor terminals).

When a current source is connected to the two probes in Fig. 11.9, the current density in the material can be found using the image theorem described in Chapter 10 when we discussed grounding electrodes. The current density vector is the superposition of the current density from the point from which the current is "injected" into the material and the current density from the point out of which the current is returning to the generator (which is just the negative of the first current with respect
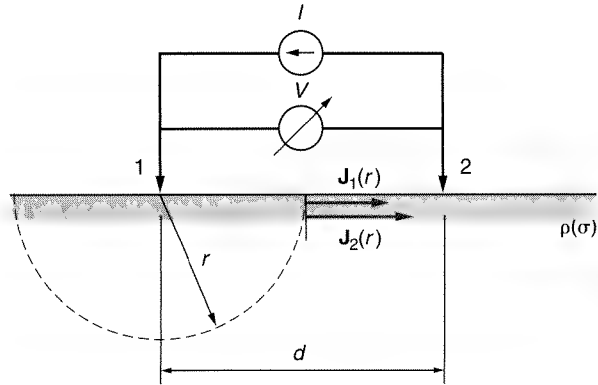
Figure 11.9 Two-point probe for resistivity measurements

to the surface of the material). As shown in the figure, this current density vector can easily be found only along the line connecting the two probes right underneath the material-air interface. Once we know the current density $\mathbf{J} = \mathbf{J}_1 + \mathbf{J}_2$, we can use Ohm's law $\mathbf{E} = \rho\mathbf{J}$ to find the expression for the voltage drop between the two probes in terms of the source current $I$ and the unknown resistivity $\rho$:

$$V = \int_{\text{probe 1}}^{\text{probe 2}} \mathbf{E} \cdot d\mathbf{l} = \int_{\text{probe 1}}^{\text{probe 2}} \rho\mathbf{J} \cdot d\mathbf{l}.$$

It is left as an exercise for the reader to write the expression for the current density and to attempt solving this integral. When solving the integral, note that the limits of integration should be the radii of the probe contacts. But because it is almost impossible to accurately know the radii of the probe contacts, we cannot accurately measure the properties of a solid material by this method.

To avoid this difficulty in measuring resistivity we use a four-point probe instead (Fig. 11.10). This time, the current is injected into the material from the outer
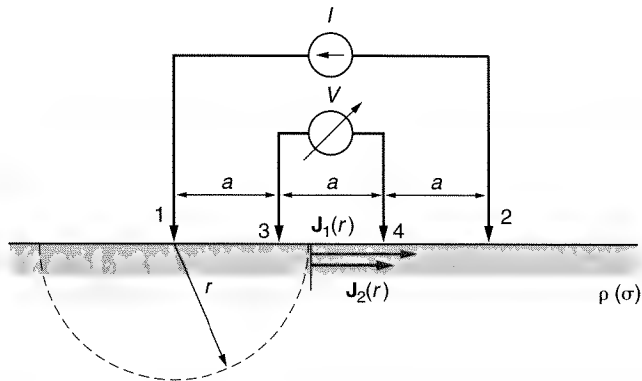


Figure 11.10 Four-point probe for resistivity measurements

two probes, but the voltage is sampled between the two inner probes. Because the two voltage probes can be pointlike, and because they are far from the current probes of radii known only approximately, this results in a much more accurate measurement of the resistivity of the material. It is left as an exercise (P11.11) to find the expression for the resistivity in this case.

*Questions and problems:*   Q11.20, P11.11 to P11.17

## 11.9   Brief Overview of Other Applications

Many other applications also involve electrostatic fields, but in this section we will briefly describe only some of them.

One of the most common electrostatic applications is coating. In the car industry several coats of paint are applied to vehicles using electrostatic coating. Every washer, dryer, and refrigerator is electrostatically coated. Even such objects as golf balls and the paper you now hold in front of you have been coated for some purpose (the paper is coated in order to have a good printing surface). The basic principle of electrostatic coating is simple: the object to be coated is charged with one polarity, and the coating material with another. The coating material is sprayed into fine particles around the object and the particles are attracted to the object by electric forces and deposited upon impact. Of course, coating machines in industry are quite sophisticated because it is important to achieve uniform coats and also to use the coating material efficiently.

Imaging technology also uses an electrostatic-based device, the charge-coupled device (CCD) camera. CCD design can even be used, for example, in making the extremely sensitive cameras used in astronomy. A CCD camera consists of a large array of MOS capacitors (Fig. 11.11). We discussed MOS capacitors in Example 8.9. Incident light creates both positive and negative charge carriers inside the $p$ semiconductor (silicon). The metal electrodes are biased positively, so that the electrons are attracted to the semiconductor-oxide interface. The number of electrons under each metal electrode is an accurate measure of the number of incident light photons. How does the camera reconstruct the image as something we can see? It needs, in effect,
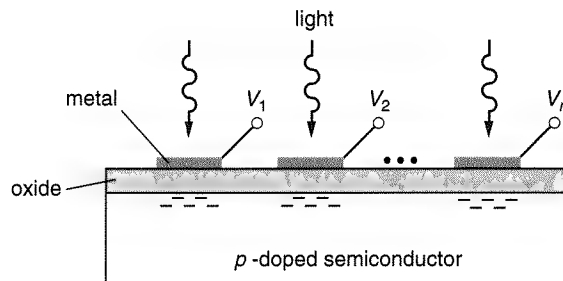
Figure 11.11 A CCD camera consists of an array of MOS capacitors
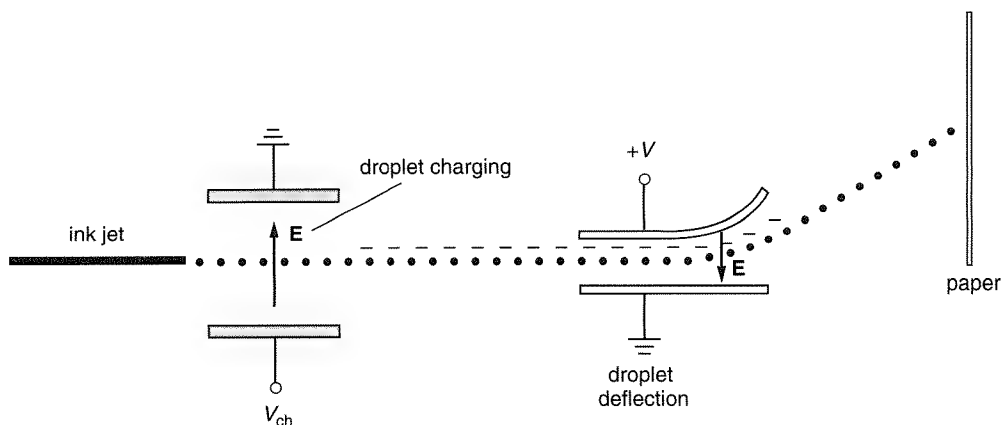
Figure 11.12 Basic components of an ink-jet printer

to measure the number of electrons in each capacitor. This is done by moving each group of electrons along a line in a serial fashion after this electron "pulsed" current is amplified in an amplifier at the end of the line. In the readout direction, each of the electrodes is biased to a progressively higher voltage, so that the electrons from the neighboring electrode are attracted to it. This transfer process can be very efficient, and in good devices only 1 out of 100,000 electrons is lost in each charge transfer step.

Another commercial application of electrostatics is nonimpact printing, for example in ink-jet printers. A basic diagram of an ink-jet printer appears in Fig. 11.12. The ink jet, on the order of 100 $\mu$m in diameter, is produced by applying pressure to an ink supply. The droplet stream, initially uncharged, passes through a cylindrical charging electrode biased at around 100 V along its axis, so that the jet and the electrode form a coaxial capacitor. The ink supply is connected to the other generator terminal, so the jet is charged by electrostatic induction. The jet is next modulated mechanically so that it turns into charged droplets, between 25 and 125 $\mu$m in diameter for 0.254-cm-high characters. The charged droplets are then deflected into the desired dot-matrix pattern by the electrostatic field between two metal plates with a voltage of 1 to 5 kV between them. The deflection in one plane is controlled by the voltage between the electrodes. The other dimension is usually controlled by the mechanical motion of the stream with respect to the paper.

Other applications of electrostatics include electrostatic motors, electrostatic generators, and electrophoresis (separation of charged colloidal particles by the electric field) used in biology (for example, to separate live yeast cells from dead ones). Electrophoresis is used in biochemistry to separate large charged molecules by placing them in an electric field. For example, in genetic research, DNA molecules of different topological forms (e.g., supercoiled and linear DNA) are separated effectively by electrophoresis, even though they are chemically identical. Hewlett-Packard has also developed an electrophoresis instrument intended for use in the drug industry. The interested reader is referred to the *Hewlett-Packard Journal*, June 1995.

*Questions and problems:*    Q11.21

# QUESTIONS

**Q11.1.** Describe the formation of a lightning stroke.

**Q11.2.** How large are currents in a lightning stroke?

**Q11.3.** According to which physical law does thunder occur?

**Q11.4.** A spherical cloud of positive charges is allowed to disperse under the influence of its own repulsive forces. Will charges follow the lines of the electric field strength vector?

**Q11.5.** A cloud of identical, charged particles is situated in a vacuum in the gravitational field of the earth. Is there an impressed electric field in addition to the electric field of the charges themselves? Explain.

**Q11.6.** Is Eq. (11.1) valid if the charge $Q$ from time to time collides with another particle?

**Q11.7.** Explain why the left-hand side in Eq. (11.2) is as it is, and not $m(v_2 - v_1)^2/2$.

**Q11.8.** Discuss the validity of Eqs. (11.3) and (11.4) if the charge $Q$ is negative.

**Q11.9.** An electron is emitted parallel to a large conducting flat plate that is uncharged. Describe qualitatively the motion of the electron.

**Q11.10.** If the voltage between the electrodes of an air-filled parallel-plate capacitor is increased so that corona starts on the plates, what will eventually happen without increasing the voltage further?

**Q11.11.** Electric charge is continually brought on the inner surface of an isolated hollow metal sphere situated in air. Explain what will happen outside the sphere.

**Q11.12.** Give a few examples of desirable and undesirable (1) corona and (2) spark discharges.

**Q11.13.** Describe how an electrostatic pollution-control filter works.

**Q11.14.** Sketch the field that results when an uncharged spherical conductive particle is brought into an originally uniform electric field.

**Q11.15.** Describe the process of making a xerographic copy.

**Q11.16.** Explain the physical meaning of the charge transfer time constant, or dielectric relaxation constant.

**Q11.17.** Derive Eq. (11.13) and solve it.

**Q11.18.** Describe the difference in the xerographic image with and without the developer plate.

**Q11.19.** Derive the equation of particle trajectory in a forming chute process for electrostatic separation.

**Q11.20.** Why is a four-point probe measurement more precise than a two-point probe measurement?

**Q11.21.** Describe how a CCD camera works.

# PROBLEMS

**P11.1.** Calculate the voltage between the two feet of a person (0.5 m apart), standing $r = 20$ m away from a 10-kA lightning stroke, if the moderately wet homogeneous soil conductivity is $10^{-3}$ S/m. Do the calculation for the two cases when the person is standing in positions $A$ and $B$ as shown in Fig. P11.1.
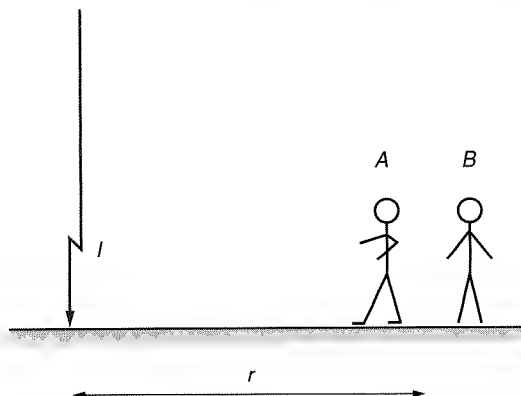
Figure P11.1 A person near a lightning stroke, standing in two positions

**P11.2.** Calculate the electric field strength above a tree that is $d = 1\,\text{km}$ away from the projection of the center of a cloud onto the earth (Fig. P11.2). Assume that because the tree is like a sharp point, the field above the tree is about 100 times that on the flat ground. As earlier, you can assume the cloud is an electric dipole above a perfectly conducting earth, with dimensions as shown in the figure, and with $Q = 4\,\text{C}$ of charge. (Note that the height of any tree is much smaller than the indicated height of the cloud.)
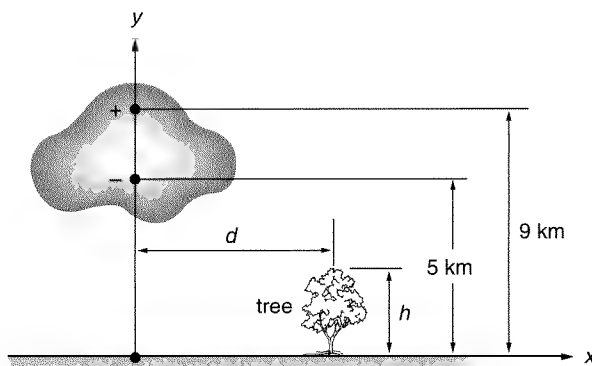


Figure P11.2 Electric field above a tree in a storm

**P11.3.** Derive the second equation in Example 11.1 from Eq. (11.1).

**P11.4.** Assuming that the initial velocity in Example 11.1 is nonzero and $x$-directed, solve for the velocity and the position of the charge $Q$ as a function of time. Plot your results.

**P11.5.** Assuming that the initial velocity in Example 11.1 is nonzero and $y$-directed, solve for the velocity and the position of the charge $Q$ as a function of time. Plot your results.

**P11.6.** A thin electron beam is formed with some convenient electrode system. The electrons in the beam are accelerated by a voltage $V_0$. The beam passes between two parallel plates, which electrostatically deflect the beam, and later falls on the screen $S$ (Fig. P11.6). Determine and plot the deflection $y_0$ of the beam as a function of the
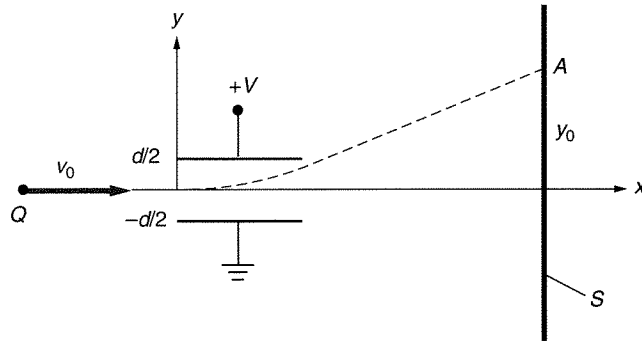
Figure P11.6 Electrostatic deflection of an electron beam

voltage $V$ between the plates. (This method is used for electrostatic deflection of the electron beam in some cathode-ray tubes.)

**P11.7.** A beam of charged particles that have positive charge $Q$, mass $m$, and different velocities enters between two closely spaced curved metal plates. The distance $d$ between the plates is much smaller than the radius $R$ of their curvature (Fig. P11.7). Determine the velocity $v_0$ of the particles that are deflected by the electric field between the plates so that they leave the plates without hitting any of them. Note that this is a kind of filter for charged particles, resulting in a beam of particles of the same velocity.
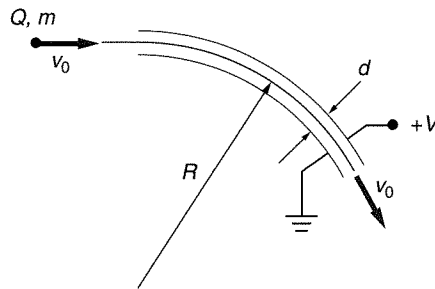


Figure P11.7 An electrostatic velocity-filter of charged particles

**P11.8.** A metal sphere is placed in a uniform electric field $E_0$. What is the maximum value of this field that does not produce air breakdown when the metal ball is brought into it?

**P11.9.** Calculate the dielectric relaxation constants for selenium, $n$-doped silicon with carrier concentration $n = 10^{16}$ cm$^{-3}$, and $n$-doped galium arsenide with concentration $n = 10^{16}$ cm$^{-3}$. For semiconductors, such as silicon and galium arsenide, the conductivity is given by $\sigma = Q\mu n$, where $Q$ is the electron charge. $\mu$ is a property of electrons inside a material, and it is called the mobility (defined as $v = \mu E$, where $v$ is the velocity of charges that moved by a field E). For silicon, $\mu = 0.135\,\text{m}^2/\text{Vs}$ and $\eta_r = 12$, and for galium arsenide, $\mu = -.86\text{m}^2/\text{Vs}$ and $\epsilon_r = 11$. for selenium, $\rho = 10^{12}$ $\Omega$-m and $\epsilon_r = 6.1$.

**P11.10.** How far do 1-mm-diameter quartz particles charged with $Q = 1\,pC$ need to fall in a field $E = 2 \cdot 10^5\,V/m$ in order to be separated by $0.5\,m$ in a forming chute separation process? The mass density of quartz is $\rho_m = 2.2\,g/cm^3$.

**P11.11.** Find the expression for determining resistivity from a four-point probe measurement, as in Fig. P11.11.
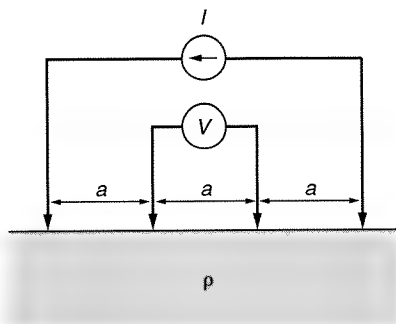


Figure P11.11 A four-point probe measurement

**P11.12.** Using the information given in P11.19, for a measured resistivity of $10\,\Omega$-cm, determine the corresponding charge concentration of (1) silicon and (2) gallium arsenide.

**P11.13.** A Wenner array used in geology is shown in Fig. P11.13. This instrument is used for determining approximately the depth of a water layer under ground. First the electrodes are placed close together, and the resistivity of soil is determined. Then the electrodes are moved farther and farther apart, until the resistivity measurement changes due to the effect of the water layer. Assuming that the top layer of soil has a very different conductivity than the water layer, what is the approximate spacing between the probes, $r$, that detects a water layer at depth $h$ under the surface? The exact analysis is complicated, so think of an approximate qualitiative solution.
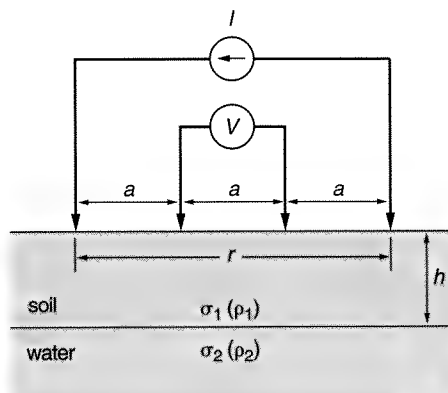


Figure P11.13 A Wenner array used in geology

**P11.14.** A thin film of resistive material is deposited on a perfect insulator. Using a four-point probe measurement, determine the expression for surface resistivity $\rho_s$ of the thin film. Assume the film is very thin.

**P11.15.** Consider an approximate circuit equivalent of a thin resistive film as in Fig. P11.15. The mesh is infinite, and all resistors are equal and have a value of $R = 1\,\Omega$. Using a two-point probe analogy, determine the resistance between any two adjacent nodes $A$ and $B$ in the mesh.
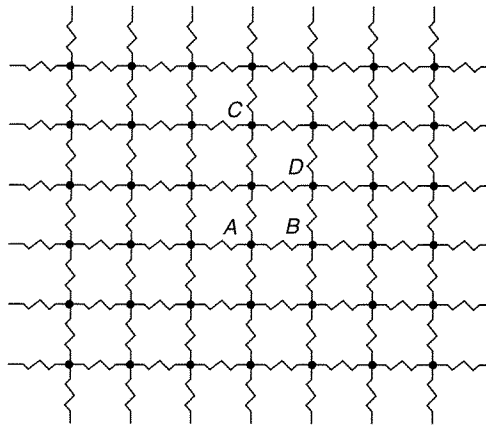


Figure P11.15 An approximate equivalent circuit of a thin resistive film

**P11.16.** Find the resistance between nodes (1) $A$ and $C$ and (2) $A$ and $D$ in Fig. P11.15.

**P11.17.** Construct an approximate equivalent circuit for a block of homogeneous resistive material. Determine the resistance between two adjacent nodes of the equivalent circuit.

PROBLEM P11.16 IS NOT SOLVABLE USING ELEMENTARY SYMMETRY TECHNIQUES. THE GENERAL SOLUTION IS IN ZEMANIAN (1991)

# 12

# Magnetic Field in a Vacuum

## 12.1 Introduction

The force between two static charges is given by Coulomb's law. Although small, this force is measurable—Coulomb established his law experimentally. If two charges are moving, it turns out that there is an additional force between them due to their motion. This force is known as the *magnetic force*.

The magnetic force between individual moving charges is extremely small when compared with the Coulomb force. Actually, it is so small that it cannot be detected experimentally if we consider just a pair of moving charges. Therefore a direct experimental derivation of the magnetic force law is not possible.

However, we can determine magnetic forces using another phenomenon—the electric current in metallic conductors—where we have an organized motion of an enormous number of electrons (practically one free electron per atom) within almost neutral substances. We can thus perform experiments in which the magnetic force can be measured with practically uncharged conductors, i.e., independently of electric forces. These experiments indicate that because of this vast number of interacting moving charges, the magnetic force between two current-carrying conductors can be much larger than the maximum obtainable electric force between them. For example, strong electromagnets can carry weights of several tons, and we know that the electric force cannot have even a fraction of that strength. Consequently, magnetic forces are used in many electrical engineering devices.

In this chapter we analyze magnetic forces in a vacuum. In many respects this chapter is of fundamental importance, just as Coulomb's law was in electrostatics. In the next chapter we will see that, in a manner somewhat analogous to that in the case of a polarized dielectric, substance in the magnetic field can be reduced to a system of electric currents situated in a vacuum.

## 12.2   Magnetic Force Between Two Current Elements

As explained, it is possible to measure the magnetic force between current-carrying conductors. In order to have a dc current in a conductor, we know that the conductor must be in the form of a closed loop, or *current loop.*

There are infinitely many different shapes and sizes of pairs of current loops for which we can measure magnetic forces. It is reasonable to assume that superposition applies to magnetic forces, as it did to electric forces. Therefore, in order to be able to determine the force between *any* two loops, we need to know the magnetic force between two arbitrarily oriented short segments of the loops. Such segments are known as *current elements.*

Consider two current loops, $C_1$ and $C_2$, with currents $I_1$ and $I_2$ (Fig. 12.1). We divide both loops into small vector line segments dl, and define a current element as the product $I$ dl (with appropriate subscripts). Note that dl is adopted to be *in the reference direction of the current along the loop.* From a large number of experimentally determined forces between various pairs of current loops, it is found that the magnetic force is always obtained correctly if it is assumed that the force between pairs of current elements is of the form

$$d\mathbf{F}_{12} = I_2 \, d\mathbf{l}_2 \times \left( \frac{\mu_0}{4\pi} \frac{I_1 \, d\mathbf{l}_1 \times \mathbf{u}_r}{r^2} \right). \tag{12.1}$$

Note that this law, like Coulomb's law, was determined experimentally. The constant of proportionality is written in the form $\mu_0/(4\pi)$ for convenience, just as in Coulomb's law it was written as $1/(4\pi\epsilon_0)$. The constant $\mu_0$ is known as the *permeability of a vacuum* (or of *free space*). It is the second of the two basic electromagnetic constants describing what we call free space or a vacuum—the first was the permittivity, $\epsilon_0$. In the SI system of units, it is defined to be *exactly*
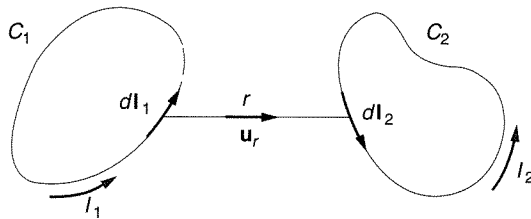


Figure 12.1 Two current loops divided into current elements

$$\mu_0 = 4\pi \cdot 10^{-7} \qquad \text{henry/m (H/m)}. \tag{12.2}$$

*(Permeability of a vacuum)*

The unit *henry per meter* (abbreviated H/m) comes from the unit for inductance, to be described in Chapter 15. Obviously, from Eq. (12.1), H/m is equal to $\text{N/A}^2$.

The force law (12.1) for two current elements is significantly more complicated than Coulomb's law. Note first that two vector cross products (see Appendix 1) are implied: we first need to determine the vector resulting from the cross product in the parentheses in Eq. (12.1) and then multiply it by the third vector in the equation. What might be confusing is that the resulting vector is oriented along the line connecting the two elements only in special cases. What might be even more perplexing is the possibility that in specific cases the magnetic force exerted by one element on another is zero, whereas the converse is not true. In all of the cases, however, the above formula has proven to be correct whenever the *total magnetic force* is calculated.

**Example 12.1—Forces between pairs of current elements.** Consider a few cases of the magnetic force law in Eq. (12.1). First let the current elements be parallel to each other and normal to the line joining them (Fig. 12.2a). From Eq. (12.1) we find that the force is repulsive if the currents in the elements are in *opposite* directions, and attractive if they are in the *same* direction. Note that this is formally opposite to the case of two charges, where like charges repel and unlike charges attract each other.

Consider now Fig. 12.2b. The force $\mathbf{F}_{12}$ in this case is zero, but the force $\mathbf{F}_{21}$ is not.

Finally, imagine we connect two parallel wires to batteries so that there is a current flowing through each wire (Fig. 12.3). We expect the two wires to exert magnetic forces on each other, according to Eq. (12.1). From the first example it is obvious that the wires attract when the currents are in the same direction and repel when they are in opposite directions.

We shall derive the expression for the force per unit length between two such wires in a later section. Just to get a feeling for the magnitude of magnetic forces, we mention one of the definitions of the *ampere*: the currents in two parallel, infinitely long wires a distance 1 m apart and situated in a vacuum are 1 A if the force on each of the wires is $2 \cdot 10^{-7}$ N per 1 m of length.
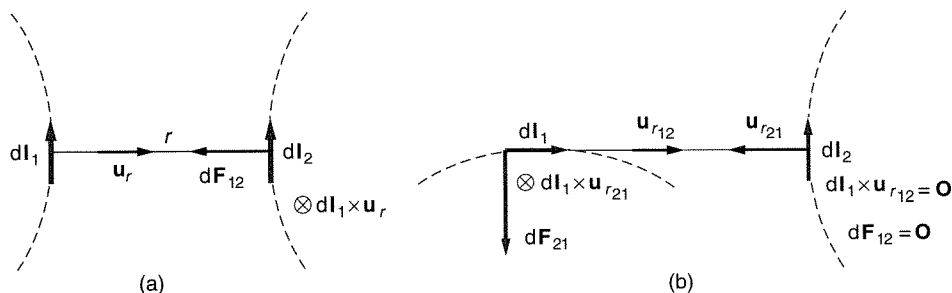


Figure 12.2 Examples of pairs of current elements
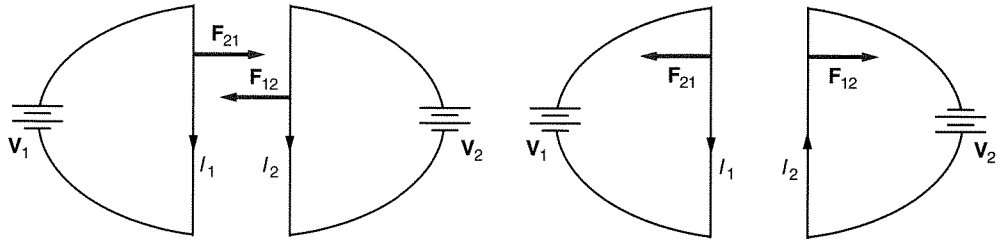
Figure 12.3 Two parallel wires can either attract or repel each other depending on the direction of the currents in them and in accordance with Eq. (12.1).

*Questions and problems:*   Q12.1 to Q12.3

## 12.3   Magnetic Flux Density and the Biot-Savart Law

We have defined the electric field as a domain of space in which there is a force on a charged particle at rest. Similarly, if in a domain of space there is a force acting on a current element or on a moving charge, we say that a *magnetic field* exists in the domain.

To characterize the magnetic field (as we did the electric field), we start from Eq. (12.1). Except for the second current element, $I_2 \, dl_2$, all the other quantities depend only on the first current element and the position and orientation of the second current element relative to the first. In analogy to the definition of the electric field intensity from Coulomb's law, we characterize the magnetic field by the *magnetic flux density vector*, d**B**, given by the term in parentheses in Eq. (12.1). Omitting the subscript "1," the expression for vector d**B** due to a current element has the form

$$dB = \frac{\mu_0}{4\pi} \frac{I \, dl \times u_r}{r^2} \qquad \text{tesla (T).} \tag{12.3}$$

*(Magnetic flux density of a current element—the Biot-Savart law)*

The unit vector $u_r$ is adopted in the same way as in the expression for the electric field intensity of a point charge: it is directed *from the source point* (i.e., the current element) *toward the field point* (i.e., the point at which we determine d**B**). This is quite similar to the Coulomb force, where we used $Q_2$ as the "test charge" and defined the rest to be a quantity characterizing the electric field of charge $Q_1$.

It is important to note that the magnetic flux density vector is perpendicular to the plane of the vectors $u_r$ and d**l**. Its orientation is determined by the right-hand rule when the vector d**l** is rotated by the shortest route toward the vector $u_r$ (see Appendix 1, section A1.2).

The magnetic flux density produced by the entire current loop $C$ is found by summing the elemental flux density vectors of all the current elements of the loop:

$$\mathbf{B} = \frac{\mu_0}{4\pi} \oint_C \frac{I \, d\mathbf{l} \times \mathbf{u}_r}{r^2} \qquad (T).$$

(12.4)

*(Magnetic flux density of a current loop—the Biot-Savart law)*

This equation and Eq. (12.3) together are referred to as the *Biot-Savart law*.

The SI unit for the magnetic flux density is the *tesla* (abbreviated T). To get a feeling for the magnitude of a tesla, let us look at a few examples. The magnetic flux density of the earth's dc magnetic field is on the order of $10^{-4}$ T. Around current-carrying conductors in a vacuum, the intensity of **B** is from about $10^{-6}$ T to about $10^{-2}$ T. In air gaps of electrical machines, the magnetic flux density can be on the order of 1 T. Superconducting magnets can produce flux densities of several dozen T. The unit T is named after the inventor Nikola Tesla, a Serbian immigrant in the United States. Tesla is responsible for ac power distribution, the first ac power plant on the Niagara Falls, the first radio remote control, induction motors, and other commonly used technologies. He was a somewhat peculiar person and his interesting life is described in many books, for example *Tesla, man out of time* by Margaret Cheney, Dorset Press (Prentice Hall), 1989.

From the definition of the magnetic flux density, it follows that the magnetic force on a current element $I \, d\mathbf{l}$ in a magnetic field of flux density **B** is given by

$$d\mathbf{F} = I \, d\mathbf{l} \times \mathbf{B} \qquad (N).$$

(12.5)

*(Magnetic force on a current element)*

Following the general definition of lines of a vector function, we define the lines of vector **B** as (generally curved) imaginary lines such that vector **B** is tangential to them at all points. For example, from Eq. (12.3) it is evident that the lines of vector **B** of a single current element are circles centered along the line of the current element and in the planes perpendicular to the element.

**Example 12.2—Magnetic flux density at the center of a circular current loop.** As an example of the application of Biot-Savart's law, let us find the magnetic flux density vector at the center of a circular loop with a dc current $I$ (Fig. 12.4). The element $d\mathbf{l}$ is as drawn in the figure. The unit vector, $\mathbf{u}_r$, is directed from the current element to the center point. The vector $d\mathbf{B}$ is pointing into the page. The intensity of the total flux density vector is given by

$$B = \oint_C dB = \frac{\mu_0}{4\pi} \oint_C \frac{I \, dl \, \sin(\pi/2)}{a^2} = \frac{\mu_0 I}{4\pi} \frac{2\pi a}{a^2} = \frac{\mu_0 I}{2a}.$$

Although the most frequent cases in practice are currents in metallic wires, which allow the use of Eq. (12.4) for determining the magnetic flux density, in some
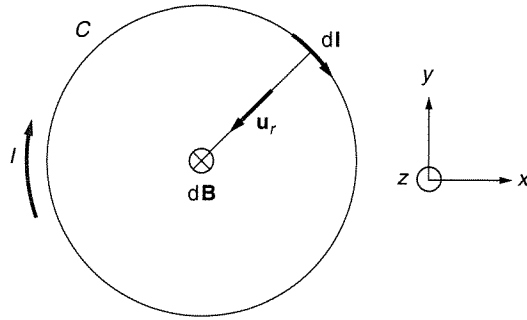
Figure 12.4 The magnetic flux density vector of a
circular current loop

cases volume currents are also encountered. We know that volume currents are described by the current density vector, $\mathbf{J}$. Let $\Delta S$ be the cross-section area of the wire. Then $I\,d\mathbf{l} = J\Delta S\,d\mathbf{l} = \mathbf{J}\,dv$ (note that $\mathbf{J}$ and $d\mathbf{l}$ have the same direction), so that the Biot-Savart law for volume currents has the form

$$\mathbf{B} = \frac{\mu_0}{4\pi} \int_v \frac{\mathbf{J} \times \mathbf{u}_r\,dv}{r^2} \qquad \text{(T)}. \tag{12.6}$$

*(Biot-Savart law for volume currents)*

At high frequencies (above about 1 MHz), currents in metallic conductors are distributed in very thin layers on conductor surfaces. These layers are so thin that they can be regarded as geometrical surfaces. It is convenient in such cases to introduce the concept of *surface current*, assuming that the current exists over the very surfaces of conductors. Of course, such a current would have infinite volume density. Therefore for its description we introduce the *surface current density*, $\mathbf{J}_s$. It is defined as current intensity, $\Delta I$, flowing "through" a line segment of length $\Delta l$ normal to the current flow, divided by the length $\Delta l$ of the segment. Evidently, the unit of surface current density is ampere per meter (A/m).

To obtain the Biot-Savart law for surface currents, let $dv = dS\,dh$, where $dh$ is the thickness of the surface current. Then $\mathbf{J}\,dv$ in the last equation becomes $\mathbf{J}\,dv = \mathbf{J}\,dS\,dh = (\mathbf{J}\,dh)\,dS = \mathbf{J}_s\,dS$, and we obtain

$$\mathbf{B} = \frac{\mu_0}{4\pi} \int_S \frac{\mathbf{J}_s \times \mathbf{u}_r\,dS}{r^2} \qquad \text{(T)}. \tag{12.7}$$

*(Biot-Savart law for surface currents)*

**Example 12.3—The magnetic force and the moment of magnetic forces on a circular loop in a uniform magnetic field.** As an application of Eq. (12.5), let us determine first the force on a circular loop of radius $a$ situated in a uniform magnetic field of flux density $\mathbf{B}$. According to Eq. (12.5), and recalling that $\mathbf{B}$ is a constant vector,

$$\mathbf{F} = \oint_C I\,d\mathbf{l} \times \mathbf{B} = I\left(\oint_C d\mathbf{l}\right) \times \mathbf{B} = 0,$$

since the line integral of dl is zero. So the magnetic force on a current loop in a uniform magnetic field is zero.

Consider now the moment of magnetic forces. Let the vector **B** be normal to the loop. It is a simple matter to conclude that the magnetic forces would then tend either to stretch the loop or to compress it (depending on the direction of **B**), but the moment on the loop is zero. Thus *only that component of the vector* **B** *that is parallel to the plane of the loop may produce the moment.*

Figure 12.5 shows the loop with the parallel vector **B**. Consider the two symmetrical elements dl and dl'. It is seen that the force d**F** = $I$ dl × **B** on element dl is directed into the page, and the force on the other element is directed out of the page. These forces tend to lift the ~~left~~ **RIGHT** half of the loop up and to press the ~~right~~ **LEFT** half of the loop down. The moment of magnetic forces on the loop (we calculate only one half and therefore multiply by two) is

$$M = 2 \int_{\alpha=0}^{\pi} I \, \mathrm{d}l \sin \alpha \, Ba \sin \alpha.$$

Noting that $\mathrm{d}l = a \, \mathrm{d}\alpha$, this becomes

$$M = 2Ia^2B \int_0^{\pi} \sin^2 \alpha \, \mathrm{d}\alpha = Ia^2 \pi B = ISB,$$

where $S$ is the area of the loop. Because the product $IS$ defines the moment of magnetic forces on the loop, it is known as the *magnetic moment of the loop* and is usually denoted by $m$. In addition, it is defined as a *vector*, with the surface $S$ determined according to the right-hand rule with respect to the current in the loop (Fig. 12.5):
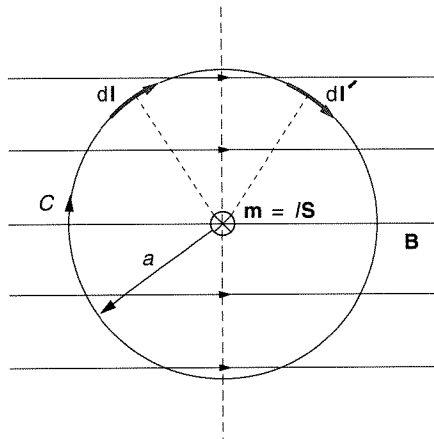


Figure 12.5 A circular current loop in a uniform magnetic field parallel to the loop plane

$$\mathbf{m} = I\mathbf{S} \quad (\text{A} \cdot \text{m}^2).$$

(12.8)

*(Definition of the magnetic moment of a current loop)*

It is a simple matter to show that the correct magnitude and direction of the moment of magnetic forces on the loop is obtained from the vector expression

$$\mathbf{M} = \mathbf{m} \times \mathbf{B} \quad (\text{N} \cdot \text{m}),$$

(12.9)

*(Moment of magnetic forces on a current loop in a uniform magnetic field)*

where **B** is the flux density vector of the uniform field *that may be in any direction*. The cross product "extracts" from **B** only the component that is parallel to the plane of the loop. Note that the moment of magnetic forces on the loop *tends to align the vectors* **m** *and* **B**.

The moment of magnetic forces is used in many applications. For example, it is used in electric motors, where current exists in a loop situated in a magnetic field.

*Questions and problems:*   Q12.4 to Q12.13, P12.1 to P12.30

## 12.4   Magnetic Flux

The term *magnetic flux* implies simply the flux of vector **B** through a surface. We shall see that it plays a very important role in magnetic circuits, and a fundamental role in one of the most important electromagnetic phenomena—electromagnetic induction.

If we have a surface $S$ in a magnetic field, the magnetic flux, $\Phi$, through $S$ is given by

$$\Phi = \int_S \mathbf{B} \cdot d\mathbf{S} \quad \text{webers (Wb)}.$$

(12.10)

*(Definition of magnetic flux)*

The unit for magnetic flux can be expressed as $\text{T} \cdot \text{m}^2$. Because of the importance of magnetic flux, this unit is given the name *weber* (abbreviated Wb). So a tesla can also be expressed as a $\text{Wb/m}^2$.

The magnetic flux has a very simple and important property—it is zero through *any* closed surface:

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \tag{12.11}$$

*(Law of conservation of magnetic flux)*

This relation is known as the *law of conservation of magnetic flux*. It follows from the expression for vector $\mathbf{B}$ of a current element in Eq. (12.3), which is proven as follows.

We know that the $\mathbf{B}$ lines of a current element are circles centered on the line that contains the current element. Along any such circle the intensity of vector $\mathbf{B}$ is constant. We can imagine the entire field of a current element divided into circular tubes formed by bunches of lines of vector $\mathbf{B}$. The magnetic flux through any such tube is the same at any cross section of the tube. Consequently, if we imagine a closed surface in this field, the magnetic flux of each of the tubes will enter the surface at one point and leave at another, making a zero contribution of the tube to the magnetic flux through the surface. Obviously, the total magnetic flux through a closed surface in the field of a current element is therefore zero.

Any distribution of current can be represented as a large number of small current elements. Because each of these elements produces zero magnetic flux through any closed surface in the field of these currents, the total flux through the closed surface is zero, i.e., Eq. (12.11) is satisfied.

An interpretation of the law of conservation of magnetic flux is that "magnetic charges" do not exist, i.e., a south and north pole of a magnet are never found separately. The law tells us also that the lines of vector $\mathbf{B}$ do not have a beginning or an end. Sometimes this last statement is phrased more loosely: it is said that the lines of vector $\mathbf{B}$ close onto themselves.

There is an important corollary of the law of conservation of magnetic flux. If we have a closed contour $C$ in the field and imagine any number of surfaces spanned over it, *the magnetic flux through any such surface, spanned over the same contour, is the same.* Just one condition needs to be fulfilled in order that this be true: the unit vector normal to all the surfaces must be normal with respect to the contour. It is customary to orient the contour and then to define the vector unit that is normal to any surface on it according to the right-hand rule (Fig. 12.6).

To prove this, consider two surfaces spanned over contour $C$, as in Fig. 12.7. They form a closed surface, to which the law of conservation of magnetic flux applies. We have, however, a specific situation concerning a vector unit normal to surface $S_2$ in the figure. If we consider $S_2$ as a part of the closed surface the vector normal should be directed outward, and if we consider it separately it should be directed according to the right-hand rule with respect to the reference contour direction, which is just opposite. So we can write

$$\oint_{S_1+S_2} \mathbf{B} \cdot d\mathbf{S} = \int_{S_1} \mathbf{B} \cdot d\mathbf{S} + \left( \int_{S_2} \mathbf{B} \cdot d\mathbf{S} \right)_{\text{outward normal}}$$

$$= \int_{S_1} \mathbf{B} \cdot d\mathbf{S} - \left( \int_{S_2} \mathbf{B} \cdot d\mathbf{S} \right)_{\text{inward normal}} = 0,$$
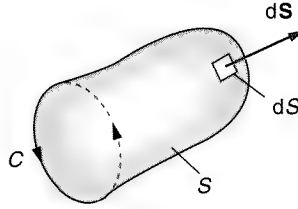
Figure 12.6 The reference direction of a vector surface element is always adopted according to the right-hand rule with respect to the reference direction of the contour defining the surface, and vice versa
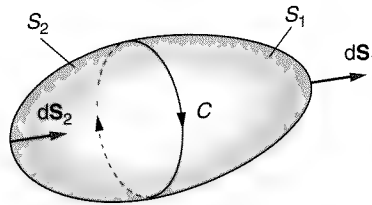


Figure 12.7 Two surfaces, $S_1$ and $S_2$, defined by a common contour, form a closed surface to which the law of conservation of magnetic flux applies

so that, indeed,

$$\int_{S_1} \mathbf{B} \cdot d\mathbf{S} = \int_{S_2} \mathbf{B} \cdot d\mathbf{S}.$$

Note that the only condition for this conclusion to be satisfied is that the flux of vector $\mathbf{B}$ is zero through any closed surface. Therefore the same conclusion holds for *any* vector that has the same property, for example the current density vector of time-invariant current fields. We shall use this conclusion in formulating Ampère's law in section 12.6.

*Questions and problems:*    Q12.14 to Q12.17

## 12.5   Electromagnetic Force on a Point Charge: The Lorentz Force

Let a current element have a cross sectional area $\Delta S$. We can then write $I\, d\mathbf{l} = J\, \Delta S\, d\mathbf{l} = NQ\mathbf{v}\, \Delta S\, dl$. Note that $\Delta S\, dl$ is the volume of the element, $N$ the number of charge carriers per unit volume, and $\mathbf{v}$ their drift velocity. Combining this result with

the expression in Eq. (12.5) for the force on a current element, we conclude that the magnetic force on a single point charge $Q$ moving at a velocity $\mathbf{v}$ in a magnetic field of (local) flux density $\mathbf{B}$ is given by

$$\mathbf{F} = Q\mathbf{v} \times \mathbf{B} \qquad (\text{N}). \qquad (12.12)$$

*(Magnetic force on a moving point charge)*

If a particle is moving both in an electric and a magnetic field, the total force (often called the *Lorentz force*) on the particle is

$$\mathbf{F} = Q\mathbf{E} + Q\mathbf{v} \times \mathbf{B} \qquad (\text{N}). \qquad (12.13)$$

*(The Lorentz force—force on a point charge moving in an electric and a magnetic field)*

**Example 12.4—The influence of B on J.** We now know that moving charges constitute a current, the current produces a magnetic field, and this magnetic field produces a force on all moving charged particles. Does this magnetic field then affect the moving charges (current) that produced it? To answer this question, we shall use the Lorentz force to examine the dependence of the current density vector $\mathbf{J}$ on both $\mathbf{E}$ and $\mathbf{B}$.

Let $N$ be the number of free charges $Q$ per unit volume at some point inside a wire carrying a current, and let the drift velocity of these charges be $\mathbf{v}$. We assume here that only one type of charge makes up the current (in metals, these are electrons). This means that $\mathbf{v} = \mathbf{J}/NQ$, so that the Lorentz force on each of the charges is

$$\mathbf{F} = Q\mathbf{E} + Q\mathbf{v} \times \mathbf{B} = Q\left(\mathbf{E} + \frac{1}{NQ}\mathbf{J} \times \mathbf{B}\right) = Q\mathbf{E}_{\text{equiv}}.$$

Let the conductivity of the conductor through which the current is flowing be $\sigma$. Then

$$\mathbf{J} = \sigma\mathbf{E}_{\text{equiv}} = \sigma\left(\mathbf{E} + \frac{1}{NQ}\mathbf{J} \times \mathbf{B}\right), \quad \text{or} \quad \frac{\mathbf{J}}{\sigma} - \mathbf{J} \times \left(\frac{\mathbf{B}}{NQ}\right) = \mathbf{E}.$$

The influence of the magnetic field on the current distribution can be neglected if the second term on the left side of this equation is much smaller than the first term. Let us compare the values of $1/\sigma$ and $|\mathbf{B}/NQ|$ that occur in practice. For example, in copper, $\sigma = 57 \cdot 10^6$ S/m, $|Q| = 1.6 \cdot 10^{-19}$ C, and $N \simeq 8.47 \cdot 10^{28}$ electrons/m$^3$. We have already said that the intensity of the magnetic flux density vector $\mathbf{B}$ rarely exceeds 1 T, so

$$|\mathbf{B}/NQ| < 7.37 \cdot 10^{-11} \quad \ll \quad 1/\sigma = 1.75 \cdot 10^{-8} \text{ m/S},$$

and we can write, approximately,

$$\mathbf{J}(\mathbf{E}, \mathbf{B}) \simeq \sigma\mathbf{E}.$$

This means that we can consider the distribution of the dc current as virtually independent of the magnetic field that it creates, i.e., that the relation $\mathbf{J} = \sigma\mathbf{E}$ is highly accurate in most cases.

*Questions and problems:*    Q12.18 to Q12.21, P12.31 to 12.33

## 12.6   Ampère's Law for Time-Invariant Currents in a Vacuum

The magnetic flux density vector **B** resulting from a time-invariant current density **J** has a very simple and important property: if we compute the line integral of **B** along any closed contour $C$, it will be equal to $\mu_0$ times the total current that flows through any surface spanned over the contour. This is *Ampère's law*:

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 \int_S \mathbf{J} \cdot d\mathbf{S}. \tag{12.14}$$

*(Ampère's law in a vacuum)*

The convention in this law is that the reference direction of the vector surface elements of $S$ is adopted according to the right-hand rule with respect to the reference direction of the contour (Fig. 12.6). In the applications of Ampère's law, it is very useful to keep in mind that the flux of the current density vector (the current intensity) is the same through all surfaces having a common boundary contour.

Ampère's law is not a new property of the magnetic field. It follows from the Biot-Savart law. Although Ampère's law itself is simple, its derivation from the Biot-Savart law is not. In addition, no physical insight is gained by this derivation, so we will not present it here. The interested reader can find it in most higher-level electromagnetics texts.

Ampère's law in Eq. (12.14) is a general law of the magnetic field of *time-invariant currents in a vacuum*. We shall see that it can be extended to cases of materials in the magnetic field, but in this form it is not valid for the magnetic field of time-varying currents.

There are two major classes of applications of Ampère's law: the determination of vector **B** in some cases with a high degree of current symmetry; and proofs of certain general properties of the magnetic field. In the examples that follow we illustrate the first kind of application of Ampère's law, similar to what we did in the case of Gauss' law.

**Example 12.5—Determination of the line integral of vector B for specified current distributions.** To learn how to use Ampère's law, consider Fig. 12.8, in which several current loops are shown. Let us determine the left-hand side in Ampère's law for the contours $C_1$, $C_2$, and $C_3$ indicated in the figure by diagonal lines.

Imagine first a surface spanned over contour $C_1$. It is necessary to assign a reference direction to this surface, using the right-hand rule with respect to the direction of $C_1$ indicated in
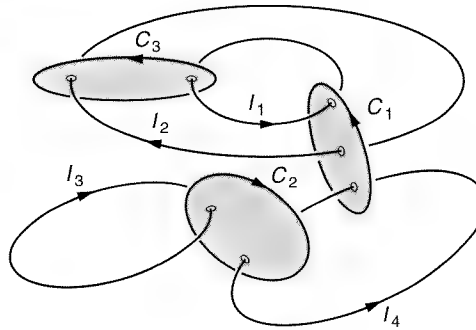
Figure 12.8 Four current loops and three contours as examples of applications of Ampère's law

Fig. 12.8. This surface is traversed in its *positive* direction by currents $I_2$ and $I_4$. It is traversed in its *negative* direction by current $I_1$. $I_3$ does not traverse it at all (or traverses it twice, once in the positive and once in the negative direction, depending on the surface we imagine). Therefore the magnetic field due to these current loops is such that the line integral of vector **B** along $C_1$ equals *exactly* $\mu_0(-I_1 + I_2 + I_4)$.

It is left as an exercise for the reader to prove that the magnetic field is also such that the line integral of **B** along $C_2$ equals $\mu_0(-I_3 - I_4)$ and along $C_3$ equals $\mu_0(I_1 - I_2)$.

**Example 12.6—Magnetic field of a straight wire of a circular cross section.** Consider now a straight, infinitely long wire of a circular cross section of radius $a$, as in Fig. 12.9. (A wire may be considered infinitely long if it is much longer than the shortest distance from it to the observation point.) There is a current of intensity $I$ in the wire distributed uniformly over its cross section, and we wish to determine vector **B** inside and outside the wire. Note that from the Biot-Savart law, the lines of the magnetic flux density vector are circles centered on the wire axis, and that the magnitude of **B** depends only on the distance $r$ from the wire axis.
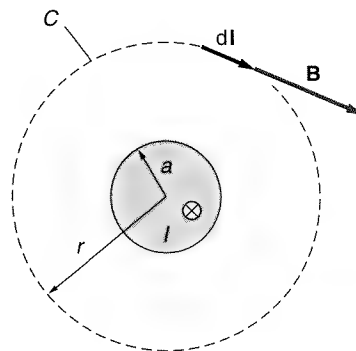


Figure 12.9 Using Ampère's law to find the magnetic flux density vector due to a current in a cylindrical wire of a circular cross section

Take a circular contour $C$ of radius $r \geq a$ centered on the wire axis. Ampère's law gives

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \oint_C B\, dl \cos 0 = B \oint_C dl = 2\pi r B = \mu_0 I,$$

so that

$$B(r) = \frac{\mu_0 I}{2\pi r} \qquad (r \geq a). \tag{12.15}$$

As long as the point is outside the wire, the radius of the wire, $a$, is irrelevant. So this expression $B$ outside a round wire is valid for any radius, even if the wire is infinitely thin.

The magnetic flux density inside the wire is obtained by applying Ampère's law to a circular contour of radius $r \leq a$. The line integral of $\mathbf{B}$ is simply $2\pi r B(r)$. The contour now *does not* encircle the entire current in the wire, but only a current $J r^2 \pi$, where $J = I/(a^2 \pi)$. The resulting magnetic flux density inside the wire is

$$B(r) = \frac{\mu_0 I r}{2\pi a^2} \qquad (r \leq a). \tag{12.16}$$

**Example 12.7—Magnetic field of a coaxial cable.** Using reasoning similar to that in Example 12.6, it is a simple matter to find the magnetic flux density due to currents $I$ and $-I$ in conductors of a coaxial cable (Fig. 12.10).

We apply Ampère's law successively to circular contours of radii $r \leq a$, $a \leq r \leq b$, $b \leq r \leq c$, and $r \geq c$. The magnetic flux density inside the inner conductor and between the conductors is the same as if the outer conductor did not exist, because the contours with radii $r \leq a$ and $a \leq r \leq b$ encircle only a part, or the total inner-conductor current. Therefore the results of the preceding example apply directly to this case, and we have

$$B(r) = \frac{\mu_0 I r}{2\pi a^2} \qquad (r \leq a). \tag{12.17}$$

$$B(r) = \frac{\mu_0 I}{2\pi r} \qquad (a \leq r \leq b). \tag{12.18}$$
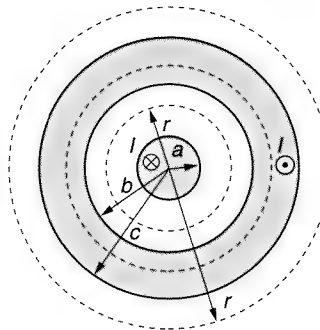


Figure 12.10 Using Ampère's law to find the magnetic flux density in a coaxial cable. The figure shows the cross section of the cable.
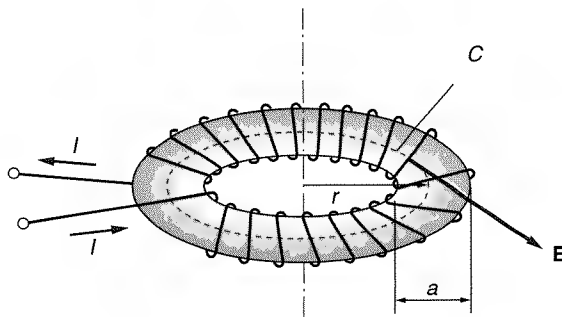
Figure 12.11 Using Ampère's law to find the magnetic flux density due to current in a toroidal coil, wound uniformly and densely with $N$ turns of thin wire

For the contour encircling the cable ($r \geq c$), the total current encircled by the contour is zero, so there is no magnetic field outside the cable. It is left for the reader as an exercise to find the expression for the magnetic flux density inside the outer cable conductor.

**Example 12.8—Magnetic field of a toroidal coil.** Consider a toroidal coil as sketched in Fig. 12.11. The cross section of the toroid is arbitrary. Assume that the coil is made of $N$ uniformly and densely wound turns with current of intensity $I$. From the Biot-Savart law, we know that the lines of vector **B** are circles centered at the toroid axis. Also, the magnitude of **B** depends only on the distance, $r$, from the axis. Applying Ampère's law yields the following expression for the magnitude, $B(r)$, of the magnetic flux density vector:

$$B = 0 \quad \text{(outside the toroid),} \tag{12.19}$$

and

$$B(r) = \frac{\mu_0 N I}{2\pi r} \quad \text{(inside the toroid).} \tag{12.20}$$

Note again that these formulas are valid for *any* shape of the toroid cross section.

As a numerical example, for $N = 1000$, $I = 2$ A, and an average toroid radius of $r = 10$ cm, we get $B = 4$ mT. This value can be larger if, for example, several layers of wire are wound on top of each other so that $N$ is larger. Alternatively, the torus can be made of a material that increases the magnetic field, to be discussed in the next chapter.

**Example 12.9—Magnetic field of a solenoid.** Assume that in the preceding example the radius $r$ of the toroid becomes very large. Then at any point inside the toroid, the toroid looks locally as if it were a cylindrical coil. Such a coil is sketched in Fig. 12.12 and is known as a solenoid. (The term *solenoid* comes from a Greek word that roughly means "tubelike.")

Outside an infinitely long solenoid the flux density vector is zero. Inside, it is given by Eq. (12.20) of the preceding example, with $r$ very large. However, the expression $N/(2\pi r)$ represents the number of turns per unit length of the toroid, i.e., of the solenoid. If we keep the number of turns per unit length constant and equal to $N'$, from Eq. (12.20) we obtain

$$B = \mu_0 N' I \quad \text{(inside a solenoid).} \tag{12.21}$$
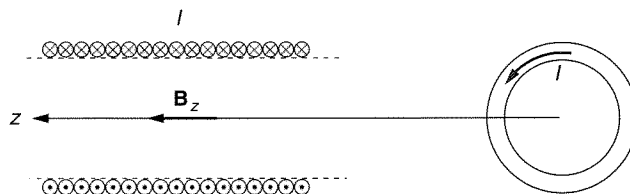
Figure 12.12 A solenoidal coil of circular cross section, wound uniformly and densely with $N'$ turns of thin wire per unit length

Note that the field inside a very long solenoid is *uniform*, and that the expression is valid for *any* cross section of the solenoid.

As a numerical example, for $N' = 2000$ windings/m, and $I = 2$ A, we get $B \simeq 5$ mT.

**Example 12.10—Magnetic field of a single current sheet and of two parallel current sheets (a strip line).** Consider a large conducting sheet with constant surface current density $\mathbf{J}_s$ at all points, as in Fig. 12.13a. From the Biot-Savart law, vector $\mathbf{B}$ is parallel to the sheet and perpendicular to vector $\mathbf{J}_s$, and $\mathbf{B}$ is directed in opposite directions on the two sides of the sheet, as indicated in the figure. What we do not know is the dependence of $B$ on $\mathscr{E}$. This can be determined using Ampère's law.    $\mathscr{y}$

Let us apply Ampère's law to the rectangular contour shown in Fig. 12.13a. Along the two rectangle sides perpendicular to the sheet, the line integral of $\mathbf{B}$ is zero because $\mathbf{B}$ is perpendicular to the line elements. Along the two sides parallel to the sheet, the integral equals $2B(x)l$. The current encircled by the contour equals $J_s l$, and Ampère's law yields

$$B(x) = \mu_0 \frac{J_s}{2} \qquad \text{(current sheet)}. \tag{12.22}$$
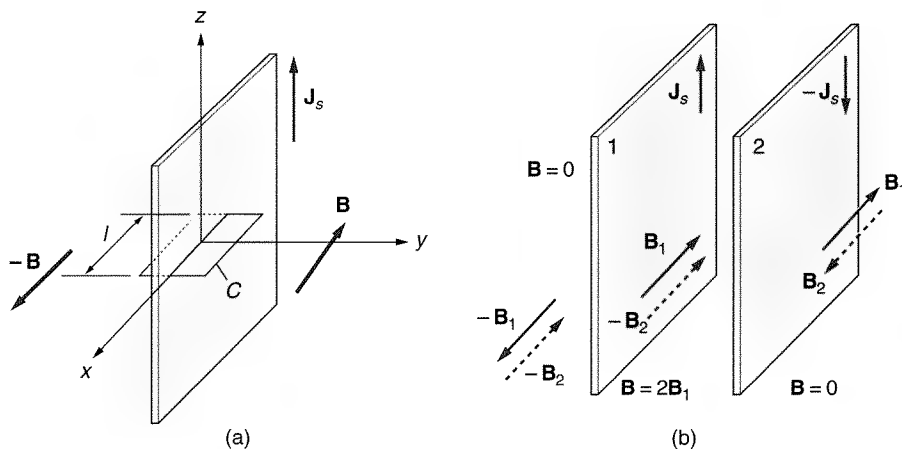


Figure 12.13 A current sheet (a) and two parallel current sheets (b)

If we have two parallel current sheets with opposite surface currents of the same magnitude (Fig. 12.13b), from the last equation and using superposition we easily find that the magnetic field outside the sheets is zero, and between the sheets

$$B = \mu_0 J_s \qquad \text{(between two parallel current sheets).} \qquad (12.23)$$

This is approximately true if the sheets are not infinite but are close to each other. Such a system is called a *strip line*.

*Questions and problems:*   Q12.22 to Q12.27, P12.34 to 12.40

# 12.7   Chapter Summary

1. Time-invariant electric currents produce a time-invariant magnetic field. This field acts with a force (the magnetic force) on a single *moving* charge or on a current element.

2. The magnetic field is described by the *magnetic flux density vector*, **B**. **B** that results from a known current distribution in a vacuum is determined from an expression known as the Biot-Savart law.

3. The total force on a moving point charge is a sum of the electric force, $Q\mathbf{E}$, and the magnetic force, $Q\mathbf{v} \times \mathbf{B}$. This sum is known as the Lorentz force.

4. A consequence of the Biot-Savart law is that the magnetic flux density vector satisfies a simple integral relation known as Ampère's law: the line integral of **B** along any closed contour in the magnetic field in a vacuum equals $\mu_0$ (the permeability of a vacuum) times the total current through any surface spanned over the contour. By convention, the direction of the vector surface elements is connected with the reference direction of the contour by the right-hand rule. Ampère's law is analogous to Gauss' law in electrostatics in that it relates the field (in this case the magnetic flux density) to the source of the field (in this case currents) through an integral equation.

## QUESTIONS

**Q12.1.** If $\mu_0$ were defined to have a different value, e.g., $\mu_0 = 1 \cdot 10^{-7}$, what would the expression for the force between two current elements be?

**Q12.2.** If we would like to have the term $I_2 \, d\mathbf{l}_2$ in Eq. (12.1) to be at the end on the right-hand side and not at the beginning, how would the expression read?

**Q12.3.** Figure Q12.3 shows four current elements (the contours they belong to are not shown). Determine the magnetic force between all possible pairs of the elements (a total of twelve expressions).
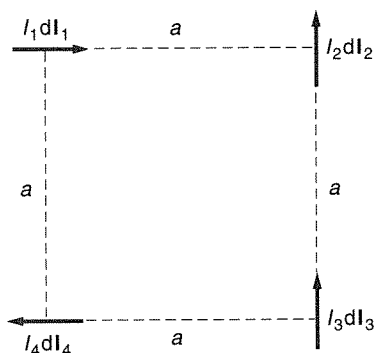
Figure Q12.3 Four current elements

**Q12.4.** What is the shape of the lines of vector **B** of a single current element? Is the magnitude of **B** constant along these lines? Is **B** constant along these lines?

**Q12.5.** Prove that $I\,dl$ for line currents is equivalent to $\mathbf{J}\,dv$ for volume currents and to $\mathbf{J}_s\,dS$ for surface currents.

**Q12.6.** Describe an approximate solution of the vector integrals in Eqs. (12.4), (12.6), and (12.7).

**Q12.7.** Assume that the lines of vector **B** converge to and are directed toward a point in space. Would that be a realistic magnetic field?

**Q12.8.** Sketch the lines of the magnetic flux density vector for two long, parallel, straight, thin conductors with equal currents when the currents are in the (1) same and (2) opposite directions.

**Q12.9.** Prove that if we have $N$ thin wire loops with currents $I$, connected in series and pressed onto one another, they can be represented as a single loop with a current $NI$. Is this conclusion valid at all points?

**Q12.10.** Starting from Eq. (12.5), write the expression for the magnetic force on a closed current loop $C$ with current $I$.

**Q12.11.** Why do we always obtain two new magnets by cutting a permanent magnet? Why do we not obtain isolated "magnetic charges"?

**Q12.12.** Knowing that the south pole–north pole direction of a compass needle aligns itself with the local direction of the vector **B**, what is the orientation of elementary current loops in the needle?

**Q12.13.** In which position is a planar current loop situated in a uniform magnetic field in stable equilibrium?

**Q12.14.** A closed surface $S$ encloses a small conducting loop with current $I$. What is the magnetic flux through $S$?

**Q12.15.** A conductor carrying a current $I$ pierces a closed surface $S$. What is the magnetic flux through $S$?

**Q12.16.** A straight conductor with a current $I$ passes through the center of a sphere of radius $R$. What is the magnetic flux through the spherical surface?

**Q12.17.** A hemispherical surface of radius $R$ is situated in a uniform magnetic field of flux density **B**. The axis of the surface makes an angle $\alpha$ with the vector **B**. Determine the magnetic flux through the surface.

**Q12.18.** Discuss the possibility of changing the kinetic energy of a charged particle by a magnetic field only.

**Q12.19.** A charge $Q$ is moving along the axis of a circular current-carrying contour normal to the plane of the contour. Discuss the influence of the magnetic field on the motion of the charge.

**Q12.20.** An electron beam passes through a region of space undeflected. Is it certain that there is no magnetic field? Explain.

**Q12.21.** An electron beam is deflected in passing through a region of space. Does this mean that there is a magnetic field in that region? Explain.

**Q12.22.** Does Ampère's law apply to a closed contour in the magnetic field of a single small charge $Q$ moving with a velocity **v**? Explain.

**Q12.23.** In a certain region of space the magnetic flux density vector **B** has the same direction at all points, but its magnitude is not constant in the direction perpendicular to its lines. Are there currents in that part of space? Explain.

**Q12.24.** An infinitely long, straight, cylindrical conductor of rectangular cross section carries a current of intensity $I$. Is it possible to determine the magnetic flux density inside and outside the conductor starting from Ampère's law? Explain.

**Q12.25.** Can the contour in Ampère's law pass through a current-carrying conductor? Explain.

**Q12.26.** What is the left-hand side in Ampère's law equal to for the five contours in Fig. Q12.26?
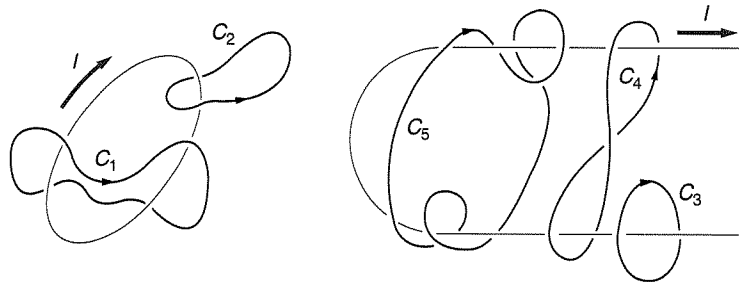


Figure Q12.26 Contours for Ampère's law

**Q12.27.** Compare Gauss' law and Ampère's law, and explain their differences and similarities.

## PROBLEMS

**P12.1.** Prove that the magnetic force on a closed wire loop of any form, situated in a uniform magnetic field, is zero.

**P12.2.** Prove that the moment of magnetic forces on a closed planar wire loop of arbitrary shape (Fig. P12.2), of area $S$ and with current $I$, situated in a uniform magnetic field

of flux density **B**, is $\mathbf{M} = \mathbf{m} \times \mathbf{B}$, where $\mathbf{m} = IS\mathbf{n}$, and **n** is the unit vector normal to $S$ determined according to the right-hand rule with respect to the direction of the current in the loop.
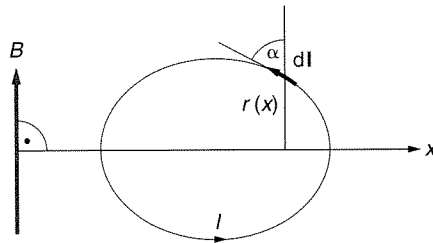


Figure P12.2 A planar current loop
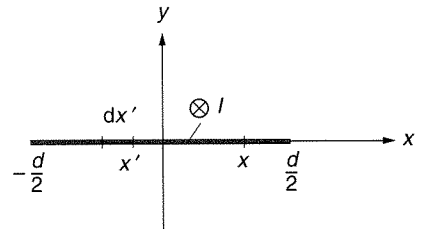


Figure P12.3 Cross section of a current strip

**P12.3.** Find the magnetic flux density vector at a point $A$ in the plane of a straight current strip (Fig. P12.3). The strip is $d$ wide and a current $I$ flows through it. Assume that point $A$ is $x$ away from the center of the strip, where $x < d/2$.

**P12.4.** A thin dielectric disk of radius $a = 10$ cm has a surface charge density of $\sigma = 2 \cdot 10^{-6}$ $C/m^2$. Find the magnetic flux density at the center of the disk if the disk is rotating at $n = 15,000$ rpm around the axis perpendicular to its surface.

**P12.5.** Determine the magnetic moment of a thin triangular loop in the form of an equilateral triangle of side $a$ with current $I$.

**P12.6.** (1) Find the magnetic flux density vector at point $P$ in the field of two very long straight wires with equal currents $I$ flowing through them. Point $P$ lies in the symmetry plane between the two wires and is $x$ away from the plane defined by the two wires. The front view of the wires is shown in Fig. P12.6a, and the top view in Fig. P12.6b. (2) What is the magnetic flux density equal to at any point in that plane if the current in one wire is $I$ and in the other $-I$?



(a)

(b)
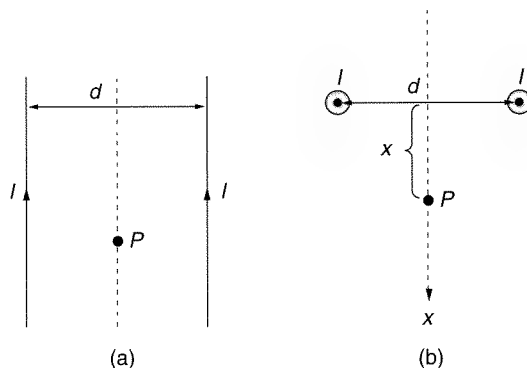
Figure P12.6 Two wires with equal currents: (a) front view, (b) top view

**P12.7.** Determine the magnetic flux density along the axis normal to the plane of a circular loop. The loop radius is $a$ and current intensity in it is $I$.

**P12.8.** We know the magnetic flux density inside a very long (theoretically infinite) thin solenoid. Usually solenoids are not long enough, so we cannot assume they are infinite. Consider a solenoid of circular cross section with a radius $a$, $b$ long, and having $N$ turns with a current $I$ flowing through them, as in Fig. P12.8. (The solenoid is actually a spiral winding, but in the figure it is shown as many closely packed circular loops.)
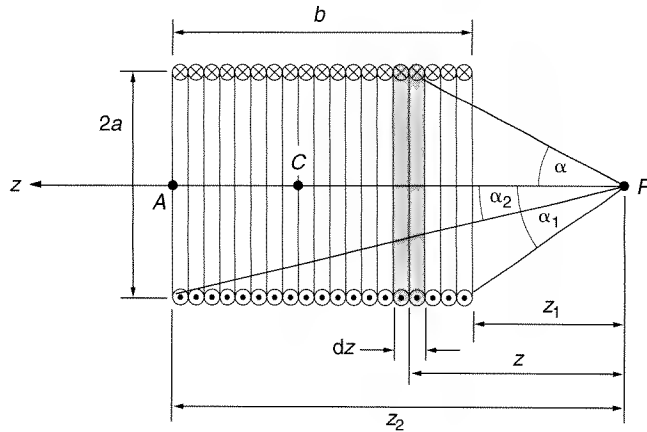


Figure P12.8 A solenoid of circular cross section

(a) How many turns are there on a length $dx$ of the coil? We can replace this small piece of the coil with a single circular loop with a current $dI$. What is $dI$ equal to?

(b) Write the expression for the magnetic flux density $d\mathbf{B}(x)$ of one of the loops with a current $dI$, at any point $P$ along the axis of the solenoid.

(c) Write the expression for the total magnetic flux density at point $P$ resulting from all of the solenoid turns. (This is an integral.)

(d) Solve the integral. It reduces to a simpler integral if you notice that $x = a/\tan\alpha$ and $a^2 + x^2 = a^2/\sin^2\alpha$.

(e) If the solenoid is thin and long, how much larger is $\mathbf{B}$ at point $C$ at the center of the solenoid than at point $A$ at the edge? Calculate the values of the magnetic flux density at these two points if $I = 2$ A, $b/a = 50$, $N = 1000$, and $b = 1$ m.

**P12.9.** A closed planar current loop carries a current of intensity $I$. Starting from the Biot-Savart law, derive the simplified integral expression for the magnitude of vector $\mathbf{B}$ for points in the plane of the loop.

**P12.10.** Derive the expression for magnitude of vector $\mathbf{B}$ of a straight current segment (Fig. P12.10). The segment is a part of a closed current loop, but only the contribution of the segment is required.
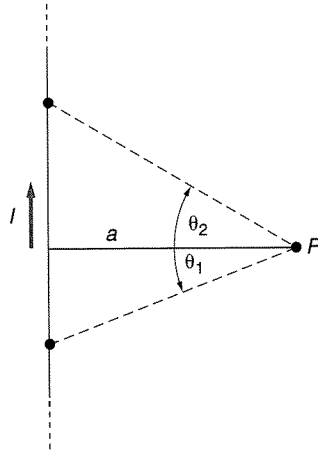
Figure P12.10 A straight current filament
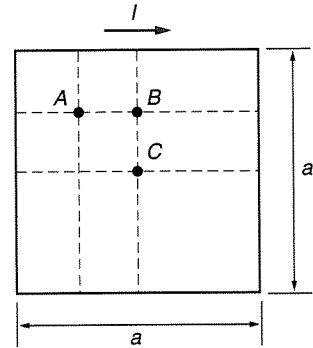
Figure P12.11 A square current loop

**P12.11.** Evaluate the magnetic flux density vector at points $A$, $B$, and $C$ in the plane of a square current loop shown in Fig. P12.11.

**P12.12.** The lengths of wires used to make a square and a circular loop with equal current are the same. Calculate the magnetic flux density at the center of both loops. In which case it is greater?

**P12.13.** Evaluate the magnetic flux density at point $A$ in Fig. P12.13.
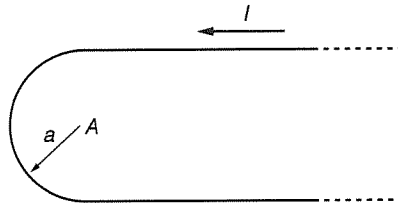


Figure P12.13 Short-circuited two-wire line

**P12.14.** Evaluate the magnetic flux density at point $A$ in the plane of a straight, flat, thin strip of width $d$ with current $I$. Assume that the point $A$ is at a distance $x$ ($x > d/2$) from the center line of the strip. Plot your result as a function of $x$.

**P12.15.** Repeat problem P12.14 for a point in a cross section of the system having coordinates $(x, y)$. Assume the origin is at the strip center line, the $x$ axis is normal to the strip, and the $y$ axis is parallel to the long side of the strip cross section. Plot the magnitude of all components of the magnetic flux density vector as a function of $x$ and $y$.

**P12.16.** A very long rectangular conductor with current $I$ has sides $a$ (along the $x$ axis) and $b$ (along the $y$ axis). Write the integral determining the magnetic flux density at any point of the $xy$ plane. Do *not* attempt to solve the integral (it is tricky).

**P12.17.** Determine and plot the magnetic flux density along the axis normal to the plane of a square loop of side $a$ carrying a current $I$.

**P12.18.** A metal spherical shell of radius $a = 10$ cm is charged with the maximum charge that does not initiate the corona on the sphere surface. It rotates about the axis passing through its center with angular velocity $\omega = 50{,}000$ rad/min. Determine the magnetic flux density at the center of the sphere.

**P12.19.** A very long, straight conductor of semicircular cross section of radius $a$ (Fig. P12.19) carries a current $I$. Determine the flux density at point $A$.
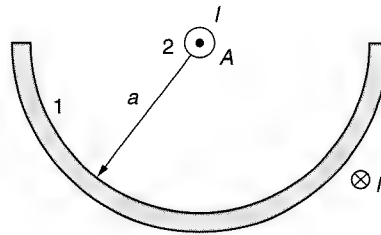


Figure P12.19 Conductor of semicircular cross section

**P12.20.** Assume in Fig. P12.19 that the thin wire 2 extends along the axis of conductor 1. Wire 2 carries the same current $I$ as conductor 1, but in the opposite direction. Determine the magnetic force per unit length on conductor 2.

**P12.21.** Determine the magnetic force on the segment $A - A'$ of the two-wire-line short circuit shown in Fig. P12.21.
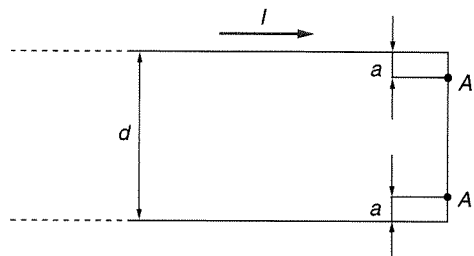


Figure P12.21 Short-circuited two-wire line

**P12.22.** Shown in Fig. P12.22 is a sketch of a permanent magnet used in loudspeakers. The lines of the magnetic flux density vector are radial, and at the position of the coil it has a magnitude $B = 1$ T. Determine the magnetic force on the coil (which is glued to the loudspeaker membrane) at the instant when the current in the coil is $I = 0.15$ A, in the indicated direction. The number of turns of the coil is $N = 10$, and its radius $a = 0.5$ cm.
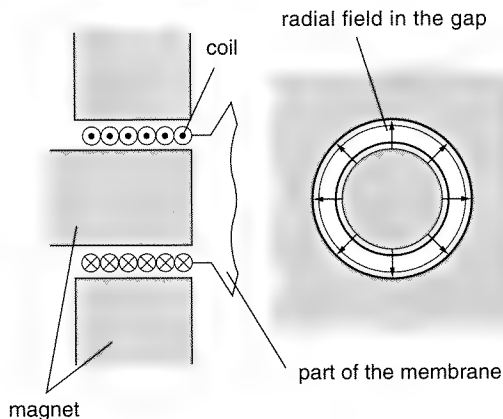
Figure P12.22 Magnet and coil of a loudspeaker

**P12.23.** Prove that the magnetic force on a *segment* of a closed current loop with current *I*, situated in a uniform magnetic field of flux density **B**, does not depend on the segment shape but only on the position of its two end points.

**P12.24.** A circular current loop of radius *a* and with current *I* is cut into halves that are in contact. It is situated in a uniform magnetic field of flux density **B** normal to the plane of the loop. (1) What should be the direction of **B** with respect to that of the current in the loop in order that the magnetic force will press the two loop halves one onto the other? (2) What is the force on each of the loop halves? Evaluate the force for *a* = 10 cm, *I* = 2 A, and *B* = 1 T. (3) What is the direction of force on the two halves of the loop due to the current in the loop itself? (Neglect this force, but note that it always exists.)

**P12.25.** Three circular loops are made of three equal pieces of wire of length *b*, one with a single turn, one with two turns, and one with three turns of wire. If the same current *I* exists in the loops and they are situated in a uniform magnetic field of flux density *B*, determine the maximal moment of magnetic forces on the three loops. Then solve for the moments if *I* = 5 A, *b* = 50 cm, and *B* = 1 T.

**P12.26.** Determine the moment of magnetic forces acting on a rectangular loop of sides *a* and *b*, and with current of intensity *I*, situated in a uniform magnetic field of flux density **B**. Side *b* of the loop is normal to the lines of **B**, and side *a* parallel to them.

**P12.27.** Two thin, parallel, coaxial circular loops of radius *a* are a distance *a* apart. Each loop carries a current *I*. Prove that at the midpoint between the loops, on their common axis, the first three derivatives of the axial magnetic flux density with respect to a coordinate along the axis are zero. (This means that the field around that point is highly uniform. Two such coils are known as *Helmholtz coils*.)

**P12.28.** Write the expression for the vector **B** inside a long circular conductor of radius *a* carrying a current *I*. To that end, use the current density vector **J** inside the conductor, and the vector **r** representing the distance of any point inside the conductor to the conductor axis.

**P12.29.** A very long cylindrical conductor of circular cross section of radius *a* has a hole of radius *b*. The axis of the hole is a distance *d* (*d* + *b* < *a*) from the conductor axis. Using

the principle of superposition and the expression for the magnetic flux density vector inside a round conductor from the preceding problem, prove that the magnetic field in the cavity is uniform.

**\*P12.30.** Prove that the divergence of the magnetic flux density vector given by the Biot-Savart law is zero.

**P12.31.** A straight, very long, thin conductor has a charge $Q'$ per unit length. It also carries a current of intensity $I$. A charge $Q$ is moving with a velocity $v$ parallel to the wire, at a distance $d$ from it, unaffected by the simultaneous action of both the electric and magnetic force. Determine the velocity $v$ of the charge, assuming the necessary correct direction of the current and the sign of the charge on the conductor.

**P12.32.** Starting from the magnetic force between two current elements, derive the expression for the magnetic force between two moving charges.

**P12.33.** Assuming that the expression for the magnetic force between two moving charges from the preceding problem is true, compare the maximal possible magnetic force between the charges with the Coulomb force between them. The charges are moving with equal velocities **v** and are at a distance $r$.

**P12.34.** A copper wire of circular cross section and radius $a = 1$ mm carries a current of $I = 50$ A. This is the largest current that can flow through the wire without damaging the conductor material. Plot the magnitude of the magnetic flux density vector as a function of distance from the center of the wire. Calculate the magnetic flux density at the surface of this wire.

**P12.35.** A plant for aluminum electrolysis uses a dc current of 15 kA flowing through a line that consists of three metal plate electrodes, as in Fig. P12.35. All the dimensions in the figure are given in centimeters. Find the approximate magnetic flux density at points $A_1$, $A_2$, and $A_3$ shown in the figure.
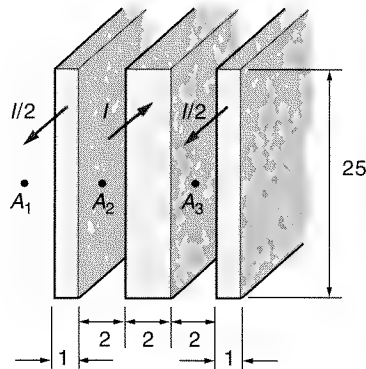


Figure P12.35 DC line for aluminum electrolysis

**P12.36.** A very long cylinder of radius $a$ has a volume charge of density $\rho$. Find the expression for the magnetic flux density vector inside as well as outside the cylinder if the cylinder is rotating around its axis with an angular velocity $\omega$. Plot your results.

**P12.37.** Find the magnetic flux density between and outside the large current sheets shown in Fig. P12.37.
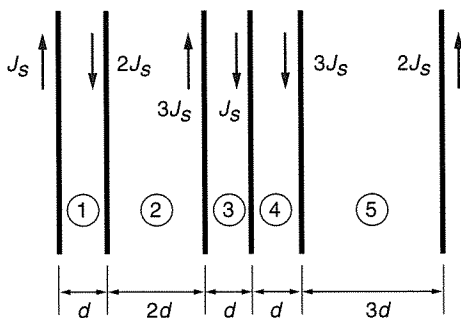


Figure P12.37 Parallel current sheets
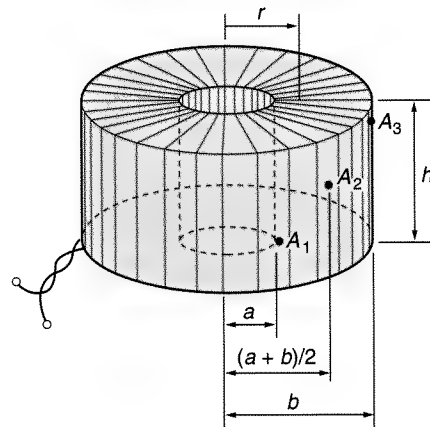


Figure P12.38 A densely wound thick toroidal coil

**P12.38.** A current $I = 0.5$ A flows through the torus winding shown in Fig. P12.38. Find the magnetic flux density at points $A_1$, $A_2$, and $A_3$ inside the torus. There are $N = 2500$ turns, $a = 5$ cm, $b = 10$ cm, and $h = 4$ cm.

**P12.39.** Find the dimensions and required number of turns for a torus such as the one in the previous problem so that the following design parameters are satisfied: (1) the magnetic flux density in the middle of the torus cross section is 6 mT; (2) the cross section of the core has dimensions $b - a = 10$ cm and $h = 10$ cm; and (3) the magnetic flux density does not vary by more than 3% from the value in the middle of the cross section. Assume you have at your disposal an insulated copper wire with a 1.5-mm diameter that can tolerate a maximum current of $I_{max} = 7.5$ A. If it is not possible to design the winding as a single-layer coil, design a multilayer winding. Note: many possible designs meet the criteria; choose the one that uses the least amount of wire, i.e., that has the lowest cost.

**P12.40.** Find and plot the magnetic flux density vector due to a current $I$ flowing through a hollow cylindrical conductor of inner radius $a$ and outer radius $b$.

# 13

# *Magnetic Fields in Materials*

## 13.1 Introduction

The effect of the electric field on materials is related simply to the existence of charges inside the atoms, not to their motion. When a body is placed in a magnetic field, however, magnetic forces act on all *moving charges* within the atoms of the material. These moving charges make the atoms and molecules inside the material look like tiny current loops. We know that the moment of magnetic forces on a current loop is such that it tends to align vectors **m** and **B**. This means that in the presence of the field, a substance becomes a large aggregate of oriented elementary current loops. These loops produce their own magnetic field, just as dipoles in a polarized dielectric produce their own electric field. Because the rest of the body does not produce any magnetic field, it is of no importance as far as the magnetic field is concerned. Therefore, a substance in the magnetic field can be visualized as a large set of oriented elementary current loops situated in a vacuum. A material in which a magnetic force has produced such oriented elementary current loops is called a *magnetized material*.

It is also possible to find macroscopic currents (not elementary loops) that produce the same magnetic field as that of all the elementary loops in a body. Therefore, it is possible to replace a material body in a magnetic field with equivalent *macroscopic* currents *situated in a vacuum*. Because we know how to determine the field of currents in a vacuum, we are able to analyze the materials in the magnetic field as well, provided that we know how to find these equivalent currents.

209

Thus we can expect certain analogies between the analysis of materials in the presence of a magnetic field and the analysis of materials in the presence of an electric field. Many of the concepts are similar, so our knowledge of the electrostatic field enables a more concise discussion of materials in the magnetic field.

## 13.2   Substances in the Presence of a Magnetic Field: Magnetization Vector

As we have already mentioned, atoms consist of a heavy positively charged nucleus and electrons that circle around the nucleus. The number of revolutions per second of an electron around the nucleus is very large—about $10^{15}$ revolutions/s. Therefore, it is reasonable to say that such a rapidly revolving electron is a small, "elementary" current loop. This picture is in fact more complicated because it turns out that electrons spin about themselves as well. However, each atom can macroscopically be viewed as a complicated system of elementary current loops. Such an elementary current loop is called an *Ampère current*. It is characterized by a *magnetic moment*, $\mathbf{m} = I\mathbf{S}$, as shown in Fig. 13.1.

Similarly to the polarization vector $\mathbf{P}$ in the case of polarized dielectrics, the *magnetization vector*, $\mathbf{M}$, describes the density of the vector magnetic moments in a magnetic material at a given point:

$$\mathbf{M} = \frac{(\sum \mathbf{m})_{\text{in } dv}}{dv} \qquad \text{(A/m)}. \tag{13.1}$$

*(Definition of magnetization vector)*

If the number of Ampère currents per unit volume at a point is $N$, and the magnetic moment of individual atoms or molecules of the substance at that point is $\mathbf{m}$, the magnetization vector can be written in the form

$$\mathbf{M} = N\mathbf{m} \qquad \text{(A/m)}. \tag{13.2}$$

*(Alternative definition of magnetization vector)*
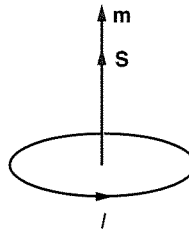


Figure 13.1 An elementary current loop is characterized by its magnetic moment, $\mathbf{m} = I\mathbf{S}$

Since the unit of $N$ is $1/m^3$, and that of **m** is $A \cdot m^2$, the unit of the magnetization vector is $A/m$.

The magnetic field of a single current loop in a vacuum can be determined from the Biot-Savart law. It can be shown that the vector **B** of such a loop at large distances from the loop is proportional to the magnetic moment of the loop, **m**. According to Eq. (13.1) we can subdivide magnetized materials into small volumes $\Delta v$, and represent such volumes (containing many Ampère currents) as a single larger Ampère current of moment **M** $\Delta v$. Consequently, if we determine the magnetization vector at all points, we can find vector **B** by *integrating* the field of these larger Ampère currents over the magnetized material. This is much simpler than adding up the fields of individual Ampère currents, since their number is prohibitively large.

*Questions and problems:*    Q13.1 to Q13.4, P13.1 and P13.2

## 13.3    Generalized Ampère's Law: Magnetic Field Intensity

We know that Ampère's law in the form in Eq. (12.14) is valid for any current distribution *in a vacuum*. We have explained, however, that a magnetized substance is just a vast number of elementary current loops in a vacuum. Therefore, we can apply Ampère's law to fields in materials, provided we find how to include these elementary currents on the right side of Eq. (12.14).

Shown in Fig. 13.2a is a surface $S$ inside a piece of magnetized material bounded by contour $C$. We know that the choice of the surface $S$ is arbitrary—the current intensity through any surface bounded by $C$ is the same. Three classes of Ampère's currents are indicated in Fig. 13.2b: those that do not pass through $S$ at all (e.g., the contour labeled 3); those that pass through $S$, but twice (contours labeled 1 and 2); and contours that encircle $C$, as the ones labeled 4 and 5. The first two types do not contribute to the total current through $S$. Contours of the third type pass through $S$ only once, and are the only ones that contribute to the total current intensity through



(a)                                    (b)
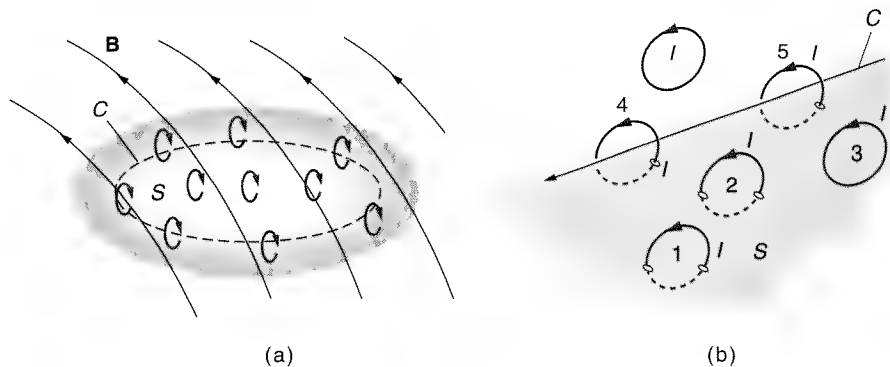
Figure 13.2 (a) A surface $S$ in a piece of magnetized material bounded by a contour $C$. (b) Enlarged section of contour $C$ and a part of surface $S$ illustrate possible relative positions of Ampère's current loops.
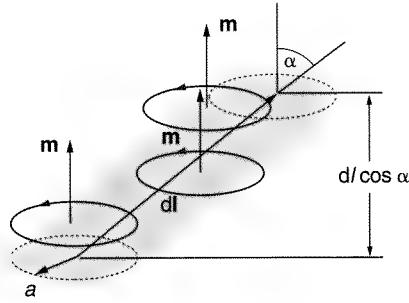
Figure 13.3 An element of contour $C$ with
neighboring current loops

$S$. What we have to find, therefore, is the total current of all the Ampère's currents
that are "strung" along $C$ like pearls in a necklace.

Consider Fig. 13.3, which shows an element dl of $C$ with neighboring current
loops. Let the radii of all the loops be $a$. It is clear that only those loops that are
centered inside an oblique cylinder of circular base $a^2\pi$ and length dl fall into the
third class of loops mentioned previously. Let the number of loops per unit volume
be $N$. The number of loops encircling dl is then $Na^2\pi\,dl\cos\alpha$, so that the total current
"strung" along dl is

$$\mathrm{d}I_{\mathrm{along\,dl}} = NIa^2\pi\,\mathrm{d}l\cos\alpha = Nm\,\mathrm{d}l\cos\alpha = M\,\mathrm{d}l\cos\alpha = \mathbf{M}\cdot\mathrm{d}\mathbf{l}. \tag{13.3}$$

Therefore, the total current "strung" along the entire contour $C$, that is, the total cur-
rent of all Ampère's currents through $S$, is given by

$$I_{\mathrm{Ampère\,through\,}S} = \oint_C \mathbf{M}\cdot\mathrm{d}\mathbf{l}. \tag{13.4}$$

It is now a simple matter to formulate Ampère's law valid for time-invariant
currents in the presence of magnetized substance:

$$\oint_C \mathbf{B}\cdot\mathrm{d}\mathbf{l} = \mu_0\left(\int_S \mathbf{J}\cdot\mathrm{d}\mathbf{S} + \oint_C \mathbf{M}\cdot\mathrm{d}\mathbf{l}\right). \tag{13.5}$$

Because the contour $C$ is the same for the integrals on the left and right side of the
equation, this can be written as

$$\oint_C (\mathbf{B}/\mu_0 - \mathbf{M})\cdot\mathrm{d}\mathbf{l} = \int_S \mathbf{J}\cdot\mathrm{d}\mathbf{S}. \tag{13.6}$$

The combined vector, $(\mathbf{B}/\mu_0 - \mathbf{M})$, has a convenient property. Its line integral
along any closed contour depends only on the *actual* current through the contour.
This is the only current we can control—switch it on and off, change its intensity or
direction, etc. For this reason the vector $(\mathbf{B}/\mu_0 - \mathbf{M})$ is defined as a new vector for the
description of the magnetic field in the presence of materials, known as the *magnetic
field intensity*, $\mathbf{H}$:

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} \qquad (A/m). \tag{13.7}$$

*(Definition of vector $\mathbf{H}$)*

With this definition, the generalized Ampère's law takes the final form

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S}. \tag{13.8}$$

*(Ampère's law for time-invariant currents in the presence of substances)*

As its special form, valid for currents in a vacuum, this form of Ampère's law is also valid *only for time-constant currents*. Also, just as its special form, it can be used for determining $\mathbf{H}$ resulting from highly symmetrical current distributions (e.g., a straight wire with a coaxial magnetic coating). As the procedure is essentially the same, the reader will find several examples of this application of Eq. (13.8) in the problems at the end of the chapter.

The definition of the magnetic field intensity vector in Eq. (13.7) is its most general definition, valid for any material. Most materials are those for which the magnetization vector, $\mathbf{M}$, is a linear function of the local vector $\mathbf{B}$ (which does the actual magnetizing of the material). According to Eq. (13.7), in such cases a linear relationship exists between any two of the three vectors, $\mathbf{H}$, $\mathbf{B}$, and $\mathbf{M}$. Usually vector $\mathbf{M}$ is expressed as

$$\mathbf{M} = \chi_m \mathbf{H} \qquad (\chi_m \text{ is dimensionless; } M \text{ is in A/m}). \tag{13.9}$$

*(Definition of magnetic susceptibility, $\chi_m$)*

The dimensionless factor $\chi_m$ is known as the *magnetic susceptibility* of the material. We then use Eq. (13.7) and express $\mathbf{B}$ in terms of $\mathbf{H}$:

$$\mathbf{B} = \mu_0(1 + \chi_m)\mathbf{H} = \mu_0 \mu_r \mathbf{H} = \mu \mathbf{H} \quad (\mu_r \text{ is dimensionless; } \mu \text{ is in H/m}). \tag{13.10}$$

*(Definition of relative permeability, $\mu_r$, and of permeability, $\mu$)*

The dimensionless factor $\mu_r = (1 + \chi_m)$ is known as the *relative permeability* of the material, and $\mu$ (H/m) as the *permeability* of the material. Materials for which Eq. (13.9) holds are *linear magnetic materials*. If it does not hold, they are *nonlinear*. If at all points of the material $\mu$ is the same, the material is said to be *homogeneous*; otherwise it is *inhomogeneous*.

Linear magnetic materials can be *diamagnetic*, for which $\chi_m < 0$ (that is, $\mu_r < 1$), or *paramagnetic*, for which $\chi_m > 0$ (that is, $\mu_r > 1$). We will discuss this in more detail

in section 13.6. Here it suffices to mention that for both diamagnetic and paramagnetic materials $\mu_r \simeq 1$, differing from unity by less than $\pm 0.001$. Therefore, in almost all applications diamagnetic and paramagnetic materials can be considered to have $\mu = \mu_0$.

Like Gauss' law, Ampère's law in Eq. (13.8) can be transformed into a differential equation, i.e., its differential form. This can easily be done by applying the so-called Stokes's theorem of vector analysis (see Appendix 1, section A1.4.7). According to this theorem, a line integral around a closed contour $C$ of a vector (e.g., vector **H**) equals the integral of the curl of that vector over *any* surface spanned over $C$:

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \nabla \times \mathbf{H} \cdot d\mathbf{S} \qquad \text{(Stokes's theorem)}.$$

So, instead of Eq. (13.8), we can write the equivalent equation

$$\int_S \nabla \times \mathbf{H} \cdot d\mathbf{S} = \int_S \mathbf{J} \cdot d\mathbf{S}. \qquad (13.11)$$

This equation must be valid for *any* contour $C$ and *any* surface spanned over it. This is possible only if the integrands of the integrals on the two sides of the equation are equal, that is,

$$\nabla \times \mathbf{H} = \mathbf{J}. \qquad (13.12)$$

*(Ampère's law in differential form, for time-invariant currents)*

This is the differential form of the generalized Ampère's law for magnetized materials and time-invariant currents. We shall use the differential form of Ampère's law in later chapters for solving various electromagnetic problems.

*Questions and problems:*   Q13.5 to Q13.8, P13.3 to P13.5

## 13.4   Macroscopic Currents Equivalent to Ampère's Currents

Let us now determine the macroscopic currents *in a vacuum* that can replace a magnetized material. We anticipate that both volume and surface currents can be expected, similar to volume and surface polarization charges in polarized dielectrics. We will see that, in analogy to homogeneous dielectrics, there are no equivalent volume currents inside *homogeneous and linear* magnetic materials with no free currents. In such cases, equivalent surface currents in a vacuum are all that is needed.

Consider a small closed contour $\Delta C$ bounding a surface of area $\Delta S$ inside a magnetized material. The total current through $\Delta C$ (i.e., through any surface bounded by $\Delta C$) is given in Eq. (13.4), where $C$ should be replaced by $\Delta C$. If we divide this by $\Delta S$, we obtain the component of the volume current density vector

due to magnetization of the material, in the direction of the normal, $\mathbf{n}$, to $\Delta S$:

$$(J_m)_{\text{normal to }\Delta S} = \lim_{\Delta S \to 0} \frac{1}{\Delta S} \oint_{\Delta C} \mathbf{M} \cdot d\mathbf{l}. \tag{13.13}$$

From mathematics we know that the right-hand side of this equation is precisely the component of $\nabla \times \mathbf{M}$ in the direction of the normal unit vector $\mathbf{n}$ (see Appendix 1, section A1.4.3). The total vector $\mathbf{J}_m$ at a point is equal to the vector sum of its three components, so that the volume density of magnetization current is given by

$$\mathbf{J}_m = \nabla \times \mathbf{M} \quad (\text{A/m}^2). \tag{13.14}$$

*(Density of magnetization currents)*

Let the material be homogeneous, of magnetic susceptibility $\chi_m$, with no macroscopic currents in it. Then

$$\mathbf{J}_m = \nabla \times \mathbf{M} = \nabla \times (\chi_m \mathbf{H}) = \chi_m \nabla \times \mathbf{H} = 0, \tag{13.15}$$

since $\nabla \times \mathbf{H} = 0$ if $\mathbf{J} = 0$, as assumed. We have thus proven that in a homogeneous magnetized material with no macroscopic currents there is no volume distribution of magnetization currents.

In addition to the preceding relationship between current density and the magnetic field vector, it can be shown that on a boundary between two magnetized materials, as in Fig. 13.4, the surface magnetic current density is given by

$$\mathbf{J}_{\text{ms}} = \mathbf{n} \times (\mathbf{M}_1 - \mathbf{M}_2) \ (\text{A/m}). \tag{13.16}$$

*(Density of magnetization surface currents)*
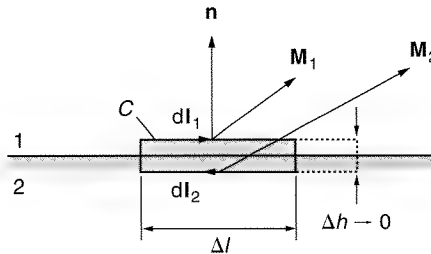


Figure 13.4 Boundary surface between two different magnetized materials

This can be proved by applying Eq. (13.4) to a small rectangular contour similar to that in Fig. 13.5̶, which is left to the reader as an exercise.

Questions and problems:   Q13.9 to Q13.14, P13.6 to P13.15

## 13.5   Boundary Conditions

Quite often it is necessary to solve magnetic problems involving inhomogeneous magnetic materials including boundaries. As in electrostatics, to be able to do this it is necessary to know the relations that must be satisfied by various magnetic quantities at two close points on the two sides of a boundary surface. We already know that all such relations are called *boundary conditions*.

We shall derive boundary conditions for the tangential components of **H** and the normal components of **B**. Assume that there are no macroscopic surface currents on the boundary surface. We use Ampère's law in Eq. (13.8) and apply it to a rectangular contour indicated in Fig. 13.5. There is no current through the contour (no macroscopic surface currents), so we find, as earlier in the case of electrostatics, that the tangential components of **H** must be equal:

$$\mathbf{H}_{1\text{tang}} = \mathbf{H}_{2\text{tang}}. \tag{13.17}$$

*(Boundary condition for tangential components of **H**, no surface currents on boundary)*

The condition for the normal components of **B** is obtained from the law of conservation of magnetic flux, Eq. (12.11). Let us apply Eq. (12.11) to the coinlike cylindrical surface with vanishingly small height shown in Fig. 13.5. In the same way as in electrostatics, where we derived the boundary condition for normal components of vector **D** from Gauss' law, we obtain
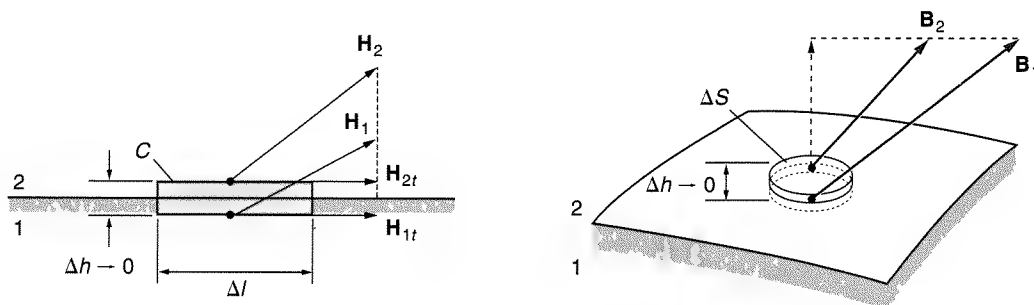


Figure 13.5 Boundary surface between two magnetic materials

$$\mathbf{B}_{1\text{norm}} = \mathbf{B}_{2\text{norm}}. \tag{13.18}$$

*(Boundary condition for normal components of* **B***)*

The boundary conditions in Eqs. (13.17) and (13.18) are valid for *any* media—linear or nonlinear. If the two media are linear, characterized by permeabilities $\mu_1$ and $\mu_2$, the two conditions can be also written in the form

$$\frac{\mathbf{B}_{1\text{tang}}}{\mu_1} = \frac{\mathbf{B}_{2\text{tang}}}{\mu_2}, \tag{13.19}$$

and

$$\mu_1 \mathbf{H}_{1\text{norm}} = \mu_2 \mathbf{H}_{2\text{norm}}. \tag{13.20}$$

**Example 13.1—Law of refraction of magnetic field lines.** If two media divided by a boundary surface are linear, the lines of vector **B** or **H** refract on the surface following a simple rule. This rule is obtained from boundary conditions in Eqs. (13.19) and (13.20).

As seen from Fig. 13.6,

$$\frac{\tan \alpha_1}{\tan \alpha_2} = \frac{H_{1\text{tang}}/H_{1\text{norm}}}{H_{2\text{tang}}/H_{2\text{norm}}} = \frac{H_{2\text{norm}}}{H_{1\text{norm}}},$$

since tangential components of **H** are equal. Using now the condition in Eq. (13.20), we obtain

$$\frac{\tan \alpha_1}{\tan \alpha_2} = \frac{\mu_1}{\mu_2}. \tag{13.21}$$

**Example 13.2—Refraction of magnetic field lines on a boundary surface between air and a material of high permeability.** The most interesting practical case of refraction of magnetic field lines is on the boundary surface between air and a medium of high permeability.
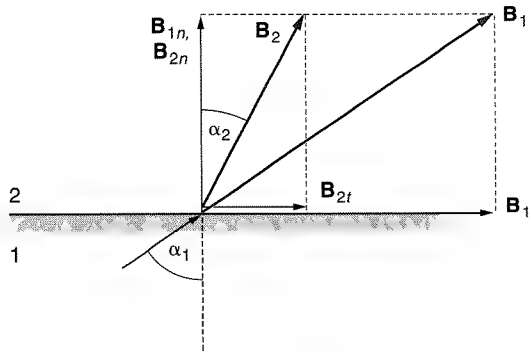


Figure 13.6 Lines of vector **B** or vector **H** refract according to Eq. (13.21)

Let air be medium 1. Then the right-hand side of Eq. (13.21) is very small. This means that $\tan\alpha_1$ must also be very small for *any* $\alpha_2$ (except if $\alpha_2 = \pi/2$, that is, if the magnetic field lines in the medium of high permeability are tangential to the boundary surface). Since for small angles $\tan\alpha_1 \simeq \alpha_1$, the *magnetic field lines in air are practically normal to the surface of high permeability*. This conclusion is very important in the analysis of electrical machines with cores of high permeability.

*Questions and problems:*    P13.16 to P13.18

## 13.6    Basic Magnetic Properties of Materials

In the absence of an external magnetic field, atoms and molecules of many materials have no magnetic moment. This is the first type of materials we will consider. The atoms and molecules of the second type of materials do have a magnetic moment, but with no external magnetic field these moments are distributed randomly, and no macroscopic magnetic field results.

### 13.6.1    DIAMAGNETIC AND PARAMAGNETIC MATERIALS

Materials of the first type are *diamagnetic materials*. When they are brought into a magnetic field, a current is induced in each atom and has the effect of *reducing* the field. (This effect is due to electromagnetic induction, to be studied in the next chapter, and exists in *all* materials. It is very small in magnitude, and in materials that are not diamagnetic it is overwhelmed by stronger effects.) Since their presence slightly *reduces* the magnetic field, diamagnetics evidently have a permeability slightly *smaller* than $\mu_0$. Examples are water ($\mu_r = 0.9999912$), bismuth ($\mu_r = 0.99984$), and silver ($\mu_r = 0.999975$).

One class of materials of the second type is *paramagnetic* materials. With no external field present, the atoms in a paramagnetic material have their magnetic moments, but these moments are oriented randomly. When a field is applied, the Ampère currents of atoms align themselves with the field to some extent. This alignment is opposed by the thermal motion of the atoms, so alignment increases as the temperature decreases and as the applied magnetic field becomes stronger. The result of the alignment of the Ampère currents is a very small magnetic field *added* to the external field. For paramagnetic materials, therefore, $\mu$ is slightly greater than $\mu_0$, and $\mu_r$ is slightly greater than one. Examples are air ($\mu_r = 1.00000036$) and aluminum ($\mu_r = 1.000021$).

The words *diamagnetic* and *paramagnetic* come from the first experiments performed to determine the magnetic nature of these materials. If a rod of a diamagnetic material is placed in a magnetic field, the magnetic moments of the atoms will try to oppose the field, and the rod will orient itself perpendicular to the lines of the magnetic field vector. The word *dia* in Greek means "across." In paramagnetics, the magnetic field of the atoms will tend to align with the external field, and the rod will orient itself parallel to the field lines. The word *para* in Greek means "along."

## 13.6.2  FERROMAGNETIC MATERIALS

The most important magnetic materials in electrical engineering are known as *ferromagnetic materials*. The name comes from the Latin word for iron, *ferrum*. They are actually paramagnetic materials, but with very strong interactions between the atoms (or molecules). As a result of these interactions, groups of atoms ($10^{12}$ to $10^{15}$ atoms in a group) are formed inside the ferromagnetic material, and in these groups the magnetic moments of all the atoms are oriented in the same direction. These groups of molecules are called *Weiss' domains*. Each domain is, in fact, a small saturated magnet. Sketches of the magnetic moments in paramagnetics and ferromagnetics as shown in Fig. 13.7.

The size of a domain varies from material to material. In iron, for example, under normal conditions the linear dimensions of the domains are $10^{-5}$ m. In some cases they can get as large as a few millimeters, or even a few centimeters across. If a piece of a highly polished ferromagnetic material is covered with fine ferromagnetic powder, it is possible to see the outlines of the domains under a microscope. The boundary between two domains is not abrupt, and it is called a *Bloch wall*. This is a region $10^{-8}$ to $10^{-6}\,\mu$m in width (500 to 5000 interatomic distances), in which the orientation of the atomic magnetic moments changes gradually.

The explanation of how and why domains form is beyond classical physics, and was quantum-mechanically described by Heisenberg in 1928. All ferromagnetics have a *crystal* structure. The ions of the crystal lattice have a magnetic moment that mostly comes from the electron spin. Ferromagnetic materials have very strong electric forces between electrons in adjacent ions, which align the magnetic moments of the ions in microscopical volumes. These forces are equivalent to extremely large intensities of vector **B**, on the order of $10^5$ T, and they act in regions that are about $10^{-2}$ mm on the side. For comparison, the strongest magnetic induction obtained in a laboratory environment is about 30 T, and the strong nuclear magnetic resonance (NMR) magnetic fields used in medical diagnostics are about 2 to 4 T.

Above a certain temperature, called the *Curie temperature*, the thermal vibrations completely prevent the parallel alignment of the molecule magnetic moments, and ferromagnetic materials become paramagnetic. This temperature is 770°C for iron (for comparison, the melting temperature of iron is 1530°C).



(a)                                         (b)
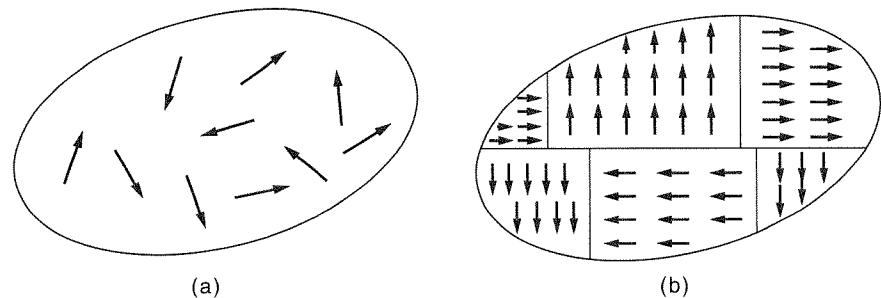
Figure 13.7 Schematic of an unmagnetized (a) paramagnetic and (b) ferromagnetic material. The arrows qualitatively show atom magnetic moments.
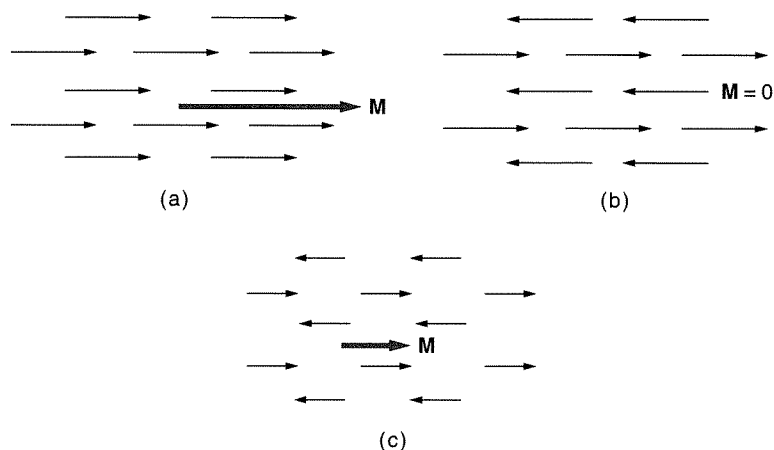
Figure 13.8 Schematic of Weiss' domains for (a) ferromagnetic, (b) antiferromagnetic, and (c) ferrite materials. The arrows represent molecular magnetic moments.

### 13.6.3   ANTIFERROMAGNETIC MATERIALS; FERRITES

Another class of materials is called *antiferromagnetics*. In these materials, the magnetic moments of adjacent molecules are antiparallel, so that the net magnetic moment is zero. (Examples are FeO, $CuCl_2$, and $FeF_2$, which are not widely used.) A subclass of antiferromagnetics called *ferrites* are widely used at radio frequencies. They also have antiparallel moments, but because of their asymmetrical structure, the net magnetic moment is not zero and Weiss' domains exist. Ferrites are weaker magnets than ferromagnetics, but they have high electrical resistivities, which makes them important for high-frequency applications, as we will see later in the text. Figure 13.8 shows a schematic comparison of the Weiss' domains for ferromagnetic, antiferromagnetic, and ferrite materials.

### 13.6.4   MAGNETIZATION CURVES OF FERROMAGNETIC MATERIALS

Ferromagnetic materials are *nonlinear*. This means that $\mathbf{B} = \mu \mathbf{H}$ is *not* true. How does a ferromagnetic material behave when placed in an external magnetic field? As the external magnetic field is increased from zero, the domains that are approximately aligned with the field increase in size. Up to a certain (not large) field magnitude, this process is reversible—if the field is turned off, the domains go back to their initial states. Above a certain field strength, the domains start rotating under the influence of magnetic forces, and this process is irreversible. The domains will keep rotating until they are all aligned with the local magnetic flux density vector. At this point, the ferromagnetic is *saturated*, and applying a stronger magnetic field does not increase the magnetization vector.
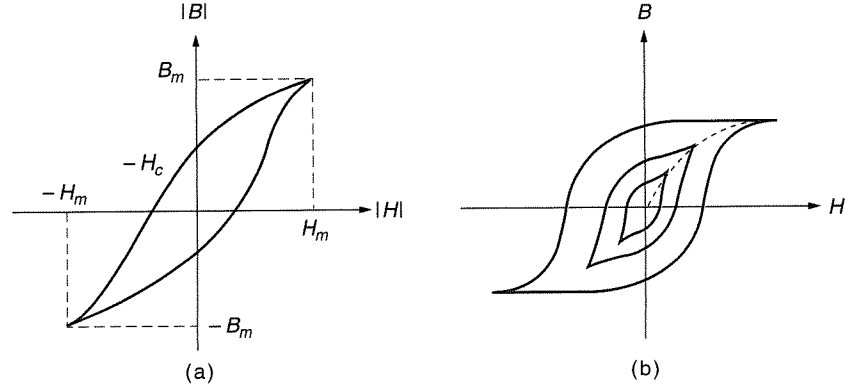
Figure 13.9 (a) Typical hysteresis loop for a ferromagnetic material. (b) The hysteresis loops for external fields of different magnitudes have different shapes. The line connecting the tips of these loops is the normal magnetization curve (shown by dashed line).

When the domains rotate there is a kind of friction between them, and this gives rise to some essential properties of ferromagnetics. If the field is turned off, the domains cannot rotate back to their original positions because they cannot overcome this friction. This means that some *permanent* magnetization is retained in the ferromagnetic material. A second consequence of friction between domains is losses to thermal energy (heat), and the third consequence is *hysteresis*, which is a term for a specific nonlinear behavior of the material. This is described by curves $B(H)$, usually measured on toroidal samples of the material. These curves are closed curves around the origin, and they are called *hysteresis loops*, shown in Fig. 13.9a. The hysteresis loops for external fields of different magnitudes have different shapes, as in Fig. 13.9b. The curve connecting the tips of these loops is known as the *normal magnetization curve*.

### 13.6.5    DEFINITIONS OF PERMEABILITY

In electrical engineering applications, the external magnetic field is usually sinusoidal. It needs to pass through several periods until the $B(H)$ curve stabilizes. The shape of the hysteresis loop depends on the frequency of the field, as well as its strength. For small field strengths it looks like an ellipse. It turns out that the ellipse approximation of the hysteresis loop is equivalent to a *complex permeability*. For sinusoidal time variation of the field, in complex notation we can write $\underline{\mathbf{B}} = \underline{\mu}\underline{\mathbf{H}}$. (This may look strange, but it is essentially the same as when we write, for example, that a complex voltage equals the product of complex impedance and complex current.) This approximation does not take saturation into account.

A ferromagnetic material in small external sinusoidal fields in complex notation can be characterized by the following parameters:

$$\underline{\sigma} = \sigma, \qquad \underline{\mu} = \mu' - j\mu'', \qquad \underline{\epsilon} = \epsilon' - j\epsilon''. \qquad (13.22)$$

Here, $\sigma$ is the conductivity of the material, and describes Joule's losses. The imaginary part, $\mu''$, of complex permeability can be shown to describe hysteresis losses in the ferromagnetic material. If $\sigma$ is small (e.g., in ferrites), in some cases it is necessary to also describe the ferromagnetic material by a complex permittivity $\underline{\epsilon}$, having the same meaning as complex permeability but with respect to a sinusoidally time-varying *electric* field. If $\sigma$ is not small, it is not necessary to characterize the material by the complex permittivity.

As explained, ferrites in small sinusoidal fields need to be described by a complex permeability and a complex permittivity. They are sometimes referred to as *ceramic ferromagnetic materials*, as opposed to *metallic ferromagnetic materials* (iron, for example). The loss mechanism is different for the two. Metallic ferromagnetics have only hysteresis losses (we will see in a later chapter that they are proportional to frequency, $f$). In ferrites, the dielectric losses, described by $\epsilon''$, are also present (they can even be predominant), and they are proportional to $f^2$.

The ratio $B/H$ (corresponding to the permeability of linear magnetic materials) for ferromagnetic materials is not a constant. It is possible to define several "permeabilities," e.g., the one corresponding to the initial, reversible segment of the magnetization curve. This permeability is known as the *initial permeability*. The range is very large, from about $500\mu_0$ for iron to several hundreds of thousands $\mu_0$ for some alloys.

The ratio $B/H$ along the normal magnetization curve (Fig. 13.9b) is known as the *normal permeability*. If we magnetize a material with a dc field, and then add to this field a small sinusoidal field, a small hysteresis loop will be obtained that will have a certain ratio $\Delta B/\Delta H$. This ratio is known as the *differential permeability*.

### 13.6.6   MEASUREMENT OF MAGNETIZATION CURVES

The curve $B(H)$ that describes the nonlinear material is usually obtained by measurement. The way this is done is shown in Fig. 13.10a. A thin toroidal core of mean radius $R$, made of the material we want to measure, has $N$ tightly wound turns of wire, and
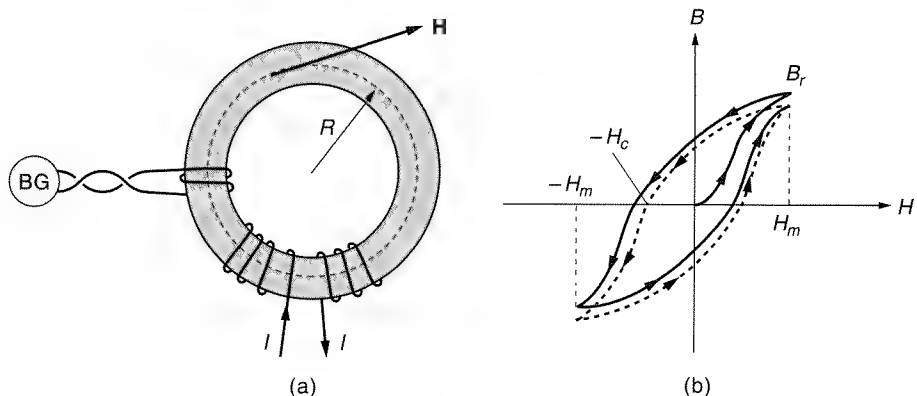


**Figure 13.10** (a) The $B(H)$ curve for a nonlinear material as measured with a ballistic galvanometer. (b) An example of measured $B(H)$ shows the formation of the hysteresis loop.

a cross-sectional area $S$. If there is a current $I$ through the winding, Ampère's law tells us that

$$H = \frac{NI}{2\pi R}. \qquad (13.23)$$

From this formula, we can calculate the magnetic field magnitude for any given current. Around the toroidal core is a second winding connected to a ballistic galvanometer. This is an instrument that measures the charge that passes through a circuit. We will see in a later chapter that the charge that flows through the circuit is proportional to the change of the magnetic flux, $\Delta Q \propto \Delta \Phi = S \Delta B$, and therefore to the change of the $B$ field as well. So by changing the current $I$ through the first winding, we can measure the curve $B(H)$ point by point. If the field $H$ is changing slowly during this process, the measured curves are called *static magnetization curves*. Figure 13.10b shows the magnetization curve measured on a previously nonmagnetized piece of material, with the *initial magnetization curve*.

*Questions and problems:*    Q13.15 to Q13.27, P13.19 to P13.22

## 13.7    Magnetic Circuits

The most frequent and important practical applications of ferromagnetic materials involve cores for transformers, motors, generators, relays, etc. The cores have different shapes, they may or may not have air gaps, and they are magnetized by a current flowing through a coil wound around a part of the core. These problems are hard to solve strictly, but the approximate analysis is accurate enough and easy because it resembles dc circuit analysis.

Consider a coil with $N$ turns and a current $I$, situated in a *linear* magnetic material. Let us look at a thin tube of small magnetic flux $\Delta\Phi$. For this case, shown in Fig. 13.11, we can write

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = NI, \qquad \Delta\Phi = B\,\Delta S, \qquad \mathbf{B} = \mu\mathbf{H}, \qquad \oint_S \mathbf{B} \cdot d\mathbf{S} = 0. \qquad (13.24)$$

The first equation from the left is Ampère's law; in the second equation both $\Delta S$ and $B$ vary along the tube; and in the last equation $S$ is any closed surface. A completely analogous set of equations can be written for dc currents in a thin current tube:

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = e, \qquad \Delta I = J\,\Delta S, \qquad \mathbf{J} = \sigma\mathbf{E}, \qquad \oint_S \mathbf{J} \cdot d\mathbf{S} = 0. \qquad (13.25)$$

In these equations, $e$ is the electromotive force in the circuit, $\Delta S$ and $J$ vary along the tube, $\sigma$ is the conductivity, and $S$ is any closed surface. Because the two sets of equations are analogous, the solutions must have the same form. For the second set of equations, Ohm's law and the two Kirchhoff's laws tell us that
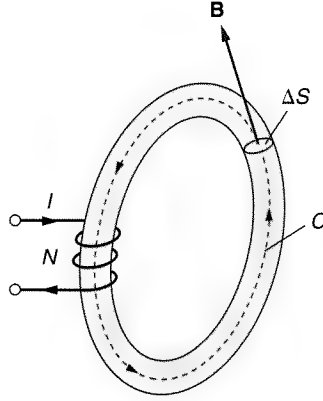
Figure 13.11 A thin magnetic circuit

$$\sum e - \sum RI = 0 \qquad \text{(any closed circuit)} \tag{13.26}$$

$$R = \int_C \frac{dl}{\sigma \, \Delta S} \qquad \left(\text{or } R = \frac{l}{\sigma S} \text{ for } \Delta S \text{ constant}\right) \tag{13.27}$$

$$\sum I = 0 \qquad \text{(any node).} \tag{13.28}$$

In a magnetic circuit, the product $NI$ plays the role of an electromotive force, and it is called the *magnetomotive force*. Permeability $\mu$ corresponds to conductivity $\sigma$. The magnetic flux corresponds to the electric current $I$. Therefore, we can write the following equations for the magnetic circuit:

$$\sum NI - \sum R_m \Phi_m = 0 \qquad \text{(any closed magnetic circuit)} \tag{13.29}$$

$$R_m = \int_C \frac{dl}{\mu \, \Delta S} \qquad \left(\text{or } R_m = \frac{l}{\mu S} \text{ for } \Delta S \text{ constant}\right) \tag{13.30}$$

$$\sum \Phi = 0 \qquad \text{(any node of the magnetic circuit).} \tag{13.31}$$

$R_m$ is called the *magnetic resistance* (or sometimes *reluctance*) of the magnetic circuit. Equations (13.29) and (13.31) are known as Kirchhoff's laws for magnetic circuits.

**Example 13.3—Thin toroidal coil as a magnetic circuit.** Let us consider a thin toroidal coil of $N$ turns, length $l$, and cross-sectional area $S$. Assume that the permeability of the core is $\mu$, and that a current $I$ is flowing through the coil. Using (13.29) and (13.30), we get

$$\Phi = \frac{NI}{R_m} = \frac{NI}{l/\mu S} = \mu \frac{NI}{l} S = \mu N'IS. \tag{13.32}$$

This is the same result we obtained when determining $B$ for the coil using Ampère's law, and using $\Phi = BS$.
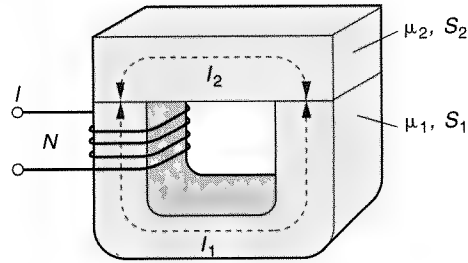
Figure 13.12 A realistic thick magnetic circuit

**Example 13.4—Thick magnetic circuits and the error using thin magnetic circuit assumption.** We have shown that the analysis of thin linear magnetic circuits is very simple. Unfortunately, real magnetic circuits are neither thin nor linear. However, analysis of thin linear magnetic circuits can be used as the basis for approximate analysis of actual magnetic circuits.

Consider a thick, U-shaped core of permeability $\mu_1$, much larger than $\mu_0$, closed by a thick bar of permeability $\mu_2$, also much larger than $\mu_0$, as shown in Fig. 13.12. $N$ turns with a current $I$ are wound on the core. The exact determination of the magnetic field in such a case is almost impossible. The first thing we can conclude is that since $\mu_1, \mu_2 \gg \mu_0$, the tangential component of the magnetic flux density $\mathbf{B}$ is much larger in the core than in the air outside it. The normal components of $\mathbf{B}$ are equal. So the magnetic flux density inside the core is generally much larger than outside. Therefore, the magnetic flux can be approximately considered to be restricted to the core. This is never exactly true, so this is the first assumption we are making.

Further, we assume that Eqs. (13.29) and (13.30) are reasonably accurate if lengths $l_1$ and $l_2$ are used as average lengths for the two circuit sections. It is interesting to show that the error in doing so is acceptable.

Consider the toroidal coil in Fig. 13.13a, the cross section of which is shown in Fig. 13.13b. The coil has $N$ densely wound turns with a current $I$, so $H = (NI)/(2\pi r)$, and $B = (\mu NI)/(2\pi r)$.



(a)                                    (b)
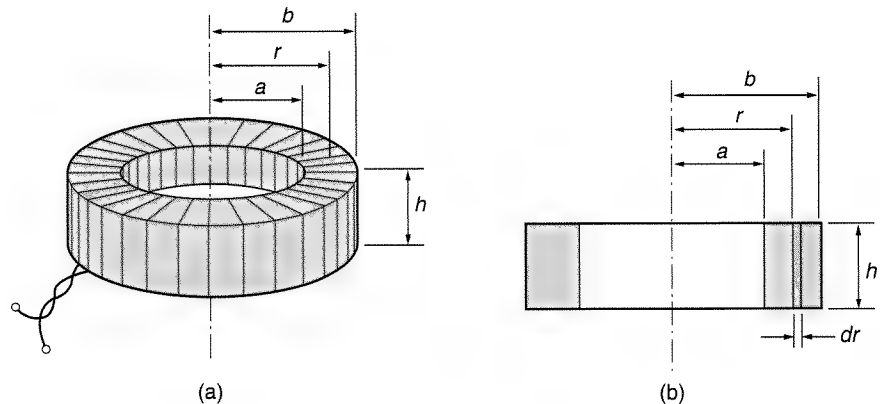
Figure 13.13 (a) A toroidal coil and (b) a cross section of the coil

The exact value of the magnetic flux through the toroid cross section is

$$\Phi_{exact} = \frac{\mu N I h}{2\pi} \int_a^b \frac{dr}{r} = \frac{\mu N I h}{2\pi} \ln\frac{b}{a}. \tag{13.33}$$

According to Eqs. (13.29) and (13.30), adopting the average length of the toroidal core, the approximate flux is

$$\Phi_{approx} = \frac{NI}{R_m} = \frac{NI}{[\pi(a+b)]/[\mu(b-a)h]} = \frac{\mu N I h}{2\pi}\frac{2(b-a)}{b+a}. \tag{13.34}$$

The relative error is

$$\frac{\Phi_{approx} - \Phi_{exact}}{\Phi_{exact}} = \frac{2(b/a-1)}{b/a\ln(b/a)} - 1, \tag{13.35}$$

which is very small even for quite thick toroids. For example, if $b/a = e = 2.718\ldots$, the error is less than 8%. So the magnetic flux in the magnetic circuit in Fig. 13.12 can be determined approximately as

$$\Phi \simeq \frac{NI}{l_1/(\mu_1 S_1) + l_2/(\mu_2 S_2)}. \tag{13.36}$$

**Example 13.5—A complex magnetic circuit.** In the case of more complicated magnetic circuits, such as the one shown in Fig. 13.14a, finding the fluxes through the branches is analogous to solving for the currents in a dc circuit. The schematic of the magnetic circuit in terms of magnetomotive forces (analogous to batteries) and magnetic resistances (analogous to resistors) is shown in Figure 13.14b. The narrow air gap is assumed not to introduce considerable flux leakage, and its reluctance is given by $R_0 = l_0/(\mu_0 S_0)$. Since $\mu_0$ is always much smaller than $\mu$, even a very narrow gap has considerable influence on the distribution of flux in the circuit



(a)                                    (b)
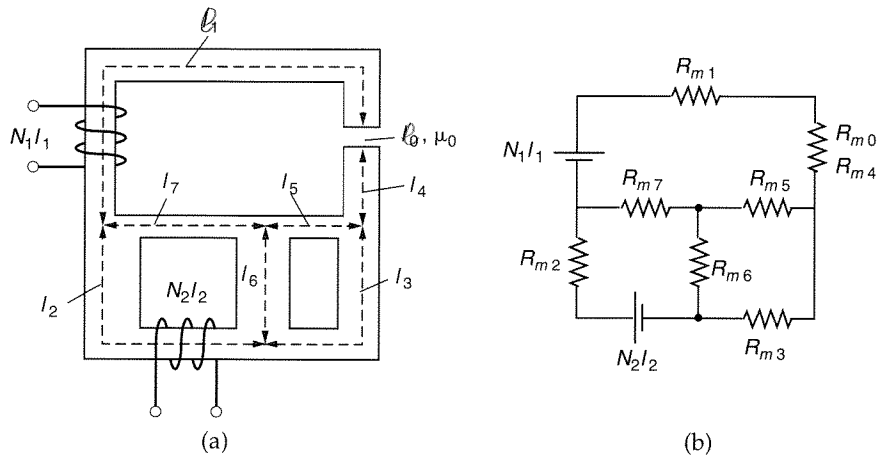
**Figure 13.14** (a) A complex magnetic circuit and (b) its representation analogous to that of an electrical network

because its reluctance is very high. The nodes in the circuit are actually surfaces enclosing the entire region where three or more branches of the circuit meet.

**Example 13.6—Analysis of a simple nonlinear magnetic circuit.** Real magnetic circuits are made of ferromagnetic materials, so they are *always nonlinear*. If the magnetization of the core is weak, the circuit can be considered as linear. This is the case for most applications of ferromagnetic materials in electronics. In power engineering, however, cores only rarely operate in the linear region. In that case, instead of using $B = \mu H$ (as in the linear analysis of magnetic circuits), we use an experimentally obtained curve $B(H)$. We shall illustrate this procedure with a simple example.

Figure 13.15a shows the initial magnetization curve of a ferromagnetic material from which the core of the circuit in Fig. 13.15b is made. We wish to determine the current intensity $I$ through the coil needed to produce a flux density in the air gap of $B \simeq 1$ T. The core was not previously magnetized. The average lengths of the core sections and the corresponding cross-sectional areas are indicated in the figure. Ignore the leakage flux.
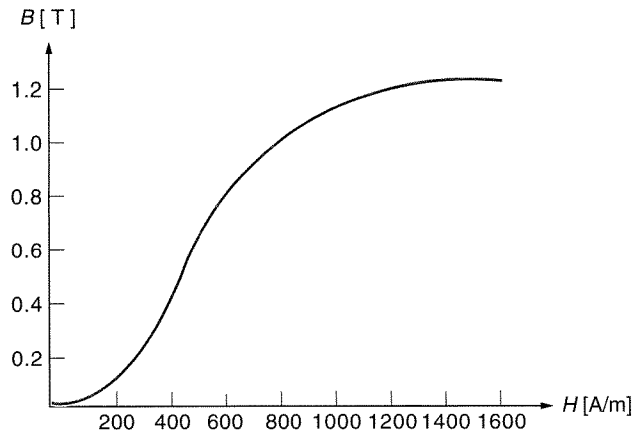


(a)
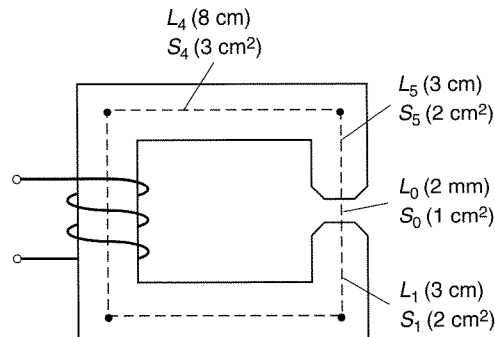


(b)

Figure 13.15 (a) Initial magnetization curve of a nonlinear material and (b) a nonlinear magnetic circuit made of the same material

TABLE 13.1 Summary of nonlinear magnetic circuit calculations for the circuit in Fig. 13.15. It can be seen from the rightmost column that $NI = 1717.8$. For all rows, the material is the nonlinear ferromagnetic, except for $k = 1$, when it is air in the gap.

| $k$ | $L_k$ (cm) | $S_k$ (cm$^2$) | $B_k$ (T) | $H_k$ (A/m) | $H_k l_k$ (A) |
|---|---|---|---|---|---|
| 0 | 0.2 | 1 | 1.0 | $8 \cdot 10^5$ | 1600 |
| 1 | 3 | 2 | 0.5 | 540 | 16.2 |
| 2 | 8 | 4 | 0.25 | 380 | 30.4 |
| 3 | 6.2 | 4 | 0.25 | 380 | 19.8 |
| 4 | 8 | 3 | 0.33 | 440 | 35.2 |
| 5 | 3 | 2 | 0.5 | 540 | 16.2 |

The current intensity producing the desired magnetic flux intensity in the air gap is determined by the following procedure:

1. From the specified $B_0$ in the air gap, the magnetic flux through the gap is $\Phi_0 = B_0 S_0$. Since we are ignoring the leakage flux, this is also the flux, $\Phi$, through the whole circuit.

2. The magnetic flux density in all parts of the circuit is determined according to $B_k = \Phi/S_k = B_0 S_0/S_k$.

3. The magnetic field intensities $H_k$ corresponding to the flux densities $B_k$ are determined next from the magnetization curve. In the air gap, $H_0 = B_0/\mu_0$.

4. The products $H_k l_k$ are calculated for all parts of the circuit.

5. Using the (approximate) expression for Ampère's law, we find the required product $NI = \sum H_k l_k$.

Table 13.1 summarizes this procedure. Note that the flux through the circuit is $\Phi = B_0 S_0 = 10^{-4}$ Wb (webers). From the table, the required intensity of the current is $I = \sum H_k l_k / N = 17.2$ A.

*Questions and problems:*   Q13.28 to Q13.33, P13.23 to P13.30

# 13.8   Chapter Summary

1. In a magnetic field, materials become magnetized, i.e., a vast number of oriented elementary current loops known as *Ampère's currents* are formed inside them. These tiny currents together produce a secondary magnetic field, which can be much larger than the field that magnetized the material.

2. With respect to their magnetic properties, all materials are divided into three basic groups: linear diamagnetic and paramagnetic materials, and nonlinear ferromagnetic materials.

3. Magnetization at a point of a magnetized material is described by the *magnetization vector*, **M**, representing the (vector) volume density of magnetic moments of Ampère's currents.

4. It is possible to generate the same magnetic field as that due to magnetization by a distribution of *macroscopic currents*, known as magnetization currents, situated in a vacuum, and derivable from the magnetization vector.

5. Ampère's law for currents in a vacuum can be extended to include magnetization effects: $\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} = \sum I$. In formulating this more general form of Ampère's law, a new vector quantity is introduced, the magnetic field intensity, $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$.

6. This extended Ampère law tells us that the line integral of $\mathbf{H}$ around any closed contour equals the total *macroscopic, controllable current* through any surface spanned over it. It is used for solving various problems, among which are problems of magnetic circuits, analogous to electric circuits.

7. The differential form of Ampère's law, $\nabla \times \mathbf{H} = \mathbf{J}$, is a partial vector differential equation in components of vector $\mathbf{H}$.

# QUESTIONS

**Q13.1.** Are any conventions implicit in the definition of the magnetic moment of a current loop? Explain.

**Q13.2.** A magnetized body is introduced into a uniform magnetic field. Is there a force on the body? Is there a moment of magnetic forces on the body? Explain.

**Q13.3.** A small body made of soft iron is placed on a table. Also on the table is a permanent magnet. If the body is pushed toward the magnet, ultimately the magnet will pull the body toward itself, so that the body will acquire a certain kinetic energy before it hits the magnet. Where did this energy come from?

**Q13.4.** Prove that the units for $\mathbf{B}$ and $\mu_0 \mathbf{M}$ are the same.

**Q13.5.** The source of the magnetic field is a permanent magnet of magnetization $\mathbf{M}$. What is the line integral of the vector $\mathbf{H}$ around a contour that passes through the magnet?

**Q13.6.** The magnetic core of a thin toroidal coil is magnetized to saturation, and then the current in the coil is switched off. The remanent (i.e., remaining) flux density in the core is $B_r$. Determine the magnetization vector and the magnetic field strength vector in the core.

**Q13.7.** Is there a magnetic field in the air around the core in question Q13.6? Explain.

**Q13.8.** Why is Eq. (13.11) valid for any contour $C$ and any surface bounded by $C$?

**Q13.9.** Why is the reference direction of the vector $\mathbf{J}_{ms}$ in Fig. 13.4 into the page?

**Q13.10.** Suppose that all atomic currents contained in the page you are reading can be oriented so that $\mathbf{m}$ is toward you. What is their macroscopic resultant, and what is (qualitatively) the magnetic field of such a "magnetic sheet"?

**Q13.11.** What are the macroscopic resultants of the microscopic currents of a short, circular, cylindrical piece of magnetized matter with uniform magnetization $\mathbf{M}$ at all points, if (1) $\mathbf{M}$ is parallel to the cylinder axis, or (2) $\mathbf{M}$ is perpendicular to the axis?

**Q13.12.** Sketch roughly the lines of vectors $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{H}$ in the two cases in question Q13.11.

**Q13.13.** A ferromagnetic cube is magnetized uniformly over its volume. The magnetization vector is perpendicular to two sides of the cube. What is this cube equivalent to in terms of the magnetic field it produces?

**Q13.14.** Is the north magnetic pole of the earth close to its geographical North Pole? Explain. (See Chapter 17.)

**Q13.15.** If a high-velocity charged elementary particle pierces a toroidal core in which there is only remanent flux density, and $\mathbf{H} = 0$, will the particle be deflected by the magnetic field? Explain.

**Q13.16.** Sketch the initial magnetization curve corresponding to a change of $H$ from zero to $-H_m$.

**Q13.17.** Suppose that the magnetization of a thin toroidal core corresponds to the point $B_r$ (remanent flux density). The coil around the core is removed, and the magnetic flux density is uniformly decreased to zero by some appropriate *mechanical or thermal treatment*. How does the point in the $B$-$H$ plane go to zero?

**Q13.18.** The initial magnetization curve of a certain ferromagnetic material is determined for a thin toroidal core. Explain the process of determining the magnetic flux in a core of the same material, but of the form shown in Fig. P13.5a.

**Q13.19.** Make a rough sketch of the curve obtained in the $B$-$H$ plane if $H$ is increased to $H_m$, then decreased to zero, then again increased to $H_m$ and decreased to zero, and so on.

**Q13.20.** Is it possible to obtain a higher remanent flux density than that obtained when saturation is attained and then $H$ is reduced to zero? Explain.

**Q13.21.** What do you expect would happen if a thin slice is cut out of a ferromagnetic toroid with remanent flux density in it? Explain.

**Q13.22.** A rod of ferromagnetic material can be magnetized in various ways. If a magnetized rod attracts a lot of ferromagnetic powder (e.g., iron filings) near its ends, and very little in its middle region, how is it magnetized?

**Q13.23.** If a small diamagnetic body is close to a strong permanent magnet, does the magnet attract or repel it? Explain.

**Q13.24.** Answer question Q13.23 for a small paramagnetic body.

**Q13.25.** While the core in question Q13.17 is still magnetized, if just one part of the core is heated above the Curie temperature, will there be a magnetic field in the air? If you think there will be, what happens when the heated part has cooled down?

**Q13.26.** Assuming that you use a large number of small current loops, explain how you can make a model of (1) a paramagnetic material and (2) a ferromagnetic material.

**Q13.27.** If the current in the coil wound around a ferromagnetic core is sinusoidal, is the magnetic flux in the core also sinusoidal? Explain.

**Q13.28.** Analyze similarities and differences for Kirchhoff's laws for dc electric circuits and magnetic circuits.

**Q13.29.** How do you determine the direction of the magnetomotive force in a magnetic circuit?

**Q13.30.** A thin magnetic circuit is made of a ferromagnetic material with an initial magnetization curve that can be approximated by the expression $B(H) = B_0 H/(H_0 + H)$, where $B_0$ and $H_0$ are constants. If the magnetic field strength, $H$, in the circuit is much smaller than $H_0$, can the circuit be considered as linear? What is in that case the permeability of the material? What is the physical meaning of the constant $B_0$?

**Q13.31.** Why can't we have a magnetic circuit with no leakage flux (stray field in the air surrounding the magnetic circuit)?

**Q13.32.** Is it possible to construct a magnetic circuit closely analogous to a dc electric circuit, if the latter is situated (1) in a vacuum, or (2) in an imperfect dielectric? Explain.

**Q13.33.** One half of the length of a thin toroidal coil is filled with a ferromagnetic material, and the other half with some paramagnetic material. Can the problem be analyzed as a magnetic circuit? Explain.

# PROBLEMS

**P13.1.** The magnetic moment of the earth is about $8 \cdot 10^{22}$ A $\cdot$ m$^2$. Imagine that there is a giant loop around the earth's equator. How large does the current in the loop have to be to result in the same magnetic moment? Would it be theoretically possible to cancel the magnetic field of the earth with such a current loop (1) on its surface, or (2) at far points? The radius of the earth is approximately 6370 km.

**P13.2.** The number of iron atoms in one cubic centimeter is approximately $8.4 \cdot 10^{22}$, and the product of $\mu_0$ and the maximum possible magnetization (corresponding to "saturation") is $\mu_0 M_{sat} = 2.15$ T. Calculate the magnetic moment of an iron atom.

**P13.3.** A thin toroid is uniformly magnetized along its length with a magnetization vector of magnitude $M$. No free currents are present. Noting that the lines of **M**, **B**, and **H** inside the toroid are circles by symmetry, determine the magnitude of **B** and prove that **H** = 0.

**P13.4.** A straight, long copper conductor of radius $a$ is covered with a layer of iron of thickness $d$. A current of intensity $I$ exists in this composite wire. Assuming that the iron permeability is $\mu$, determine the magnetic field, the magnetic flux density, and the magnetization in copper and iron parts of the wire. Note that the current density in the copper and iron parts of the wire is not the same.

**P13.5.** The ferromagnetic toroidal core sketched in Fig. P13.5a has an idealized initial magnetization curve as shown in Fig. P13.5b. Determine the magnetic field strength, the magnetic flux density, and the magnetization at all points of the core, if the core is wound uniformly with $N = 628$ turns of wire with current of intensity (1) 0.5 A, (2)
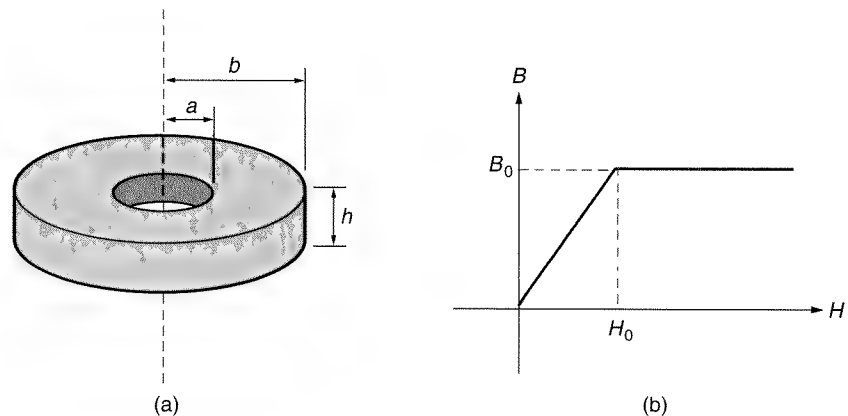


Figure P13.5 (a) A ferromagnetic core, and (b) its idealized initial magnetization curve

0.75 A, or (3) 1 A. The core dimensions are $a = 5$ cm, $b = 10$ cm, and $h = 5$ cm, and the constants of the magnetization curve are $H_0 = 1000$ A/m and $B_0 = 2$ T. For the three cases determine the magnetic flux through the core's cross section. Assume that the core was not magnetized prior to turning on the current in the winding.

**P13.6.** A straight conductor with circular cross section of radius $a$ and permeability $\mu$ carries a current $I$. A coaxial conducting tube of inner radius $b$ ($b > a$) and outer radius $c$, with no current, also has a permeability $\mu$. Determine the magnetic field intensity, magnetic flux density, and magnetization at all points. Determine the volume and surface densities of macroscopic currents equivalent to Ampère currents.

**P13.7.** Repeat problem P13.6 if the conductor and the tube are of permeability $\mu(H) = \mu_0 H / H_0$, where $H_0$ is a constant.

**P13.8.** Repeat problems P13.6 and P13.7 assuming that the tube carries a current $-I$, so that the conductor and the tube make a coaxial cable.

**P13.9.** A ferromagnetic sphere of radius $a$ is magnetized uniformly with a magnetization vector **M**. Determine the density of magnetization surface currents equivalent to the magnetized sphere.

**P13.10.** A thin circular ferromagnetic disk of radius $a = 2$ cm and thickness $d = 2$ mm is uniformly magnetized normal to its bases. The vector $\mu_0 \mathbf{M}$ is of magnitude 0.1 T. Determine the magnetic flux density vector on the disk axis normal to its bases, at a distance $r = 2$ cm from the center of the disk.

**P13.11.** A thin ferromagnetic toroidal core was magnetized to saturation, and then the current in the winding wound about the core was turned off. The remanent flux density of the core material is $B_r = 1.4$ T. Determine the surface current density on the core equivalent to the Ampère currents. If the mean radius of the core is $R = 5$ cm, and the winding has $N = 500$ turns of wire, find the current in the winding corresponding to this equivalent surface current. If the cross-sectional area of the core is $S = 1$ cm$^2$, find the magnetic flux in the core.

**P13.12.** A round ferrite rod of radius $a = 0.5$ cm and length $b = 10$ cm is magnetized uniformly over its volume. The vector $\mu_0 \mathbf{M}$ is in the direction of the rod axis, of magnitude 0.07 T. Determine the magnetic flux density at the center of one of the rod bases. Is it important whether the point is inside the rod, outside the rod, or on the very surface of the rod?

**P13.13.** Shown in Fig. P13.13 is a stripline with a ferrite dielectric. Since $a \gg d$, the magnetic field outside the strips can be neglected. Under this assumption, find the magnetic field intensity between the strips if the current in the strips is $I$. If the space between the strips is filled with a ferrite of relative permeability $\mu_r$ that can be considered
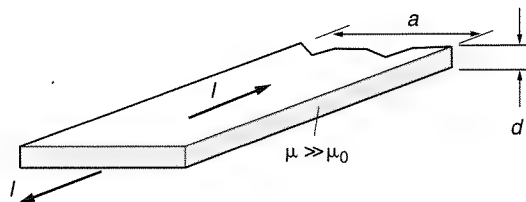


Figure P13.13 A stripline with a ferrite dielectric

constant, determine the magnetic flux density and magnetization in the ferrite, and the density of surface currents equivalent to the Ampère currents in the ferrite.

**P13.14.** The ferromagnetic cube shown in Fig. P13.14 is magnetized in the direction of the $z$ axis so that the magnitude of the magnetization vector is $M_z(x) = M_0 x/a$. Find the density of volume currents equivalent to the Ampère currents inside the cube, as well as the surface density of these currents over all cube sides. Follow the surface currents and note that in part they close through the magnetized material.
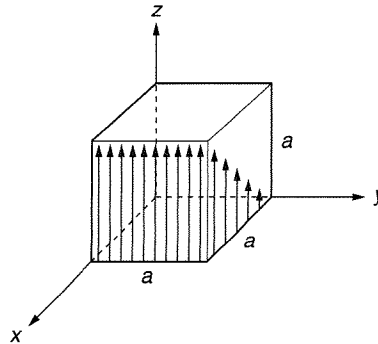


Figure P13.14 A magnetized ferromagnetic cube

**P13.15.** Prove that on the boundary surface of two magnetized materials the surface magnetization current is given by the expression $\mathbf{J}_{ms} = \mathbf{n} \times (\mathbf{M}_1 - \mathbf{M}_2)$. $\mathbf{M}_1$ and $\mathbf{M}_2$ are magnetization vectors in the two materials at close points on the two sides of the boundary, and $\mathbf{n}$ is the unit vector normal to the boundary, directed into medium 1.

**P13.16.** At a point boundary surface between air and a ferromagnetic material of permeability $\mu \gg \mu_0$ the lines of vector $\mathbf{B}$ are not normal to the boundary surface. Prove that the magnitude of the magnetic flux density vector in the ferromagnetic material is then much greater than that in the air.

**P13.17.** A current loop is in air above a ferromagnetic half-space. Prove that the field in the air due to the Ampère currents in the half-space is very nearly the same as that due to a loop below the boundary surface symmetrical to the original loop, carrying the current of the same intensity *and direction*, with the magnetic material removed. (This is the image method for ferromagnetic materials.)

**P13.18.** Inside a uniformly magnetized material of relative permeability $\mu_r$ are two cavities. One is a needlelike cavity in the direction of the vector $\mathbf{B}$. The other is a thin-disk cavity, normal to that vector. Determine the ratio of magnitudes of the magnetic flux density in the two cavities and that in the surrounding material. Using these results, estimate the greatest possible theoretical possibility of "magnetic shielding" from time-invariant external magnetic field. (We shall see that the shielding effect is greatly increased for time-varying fields.)

**P13.19.** Sketched in Fig. P13.19 is the normal magnetization curve of a ferromagnetic material. Using this diagram, estimate the relative normal and differential permeability, and plot their dependence on the magnetic field strength. What are the initial maximal permeabilities of the material?
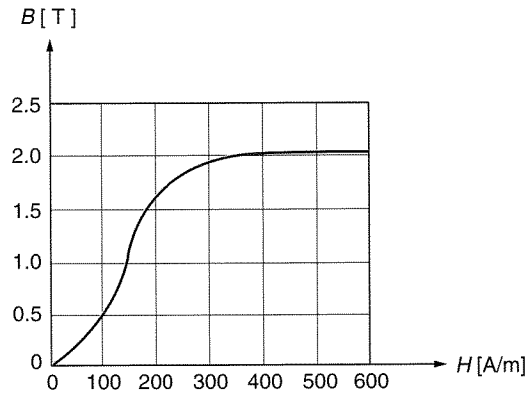
Figure P13.19 A normal magnetization curve

**P13.20.** Approximate the normal magnetization curve in Fig. P13.19 in the range $0 \leq H \leq 300\,\text{A/m}$ by a straight line segment, and estimate the largest deviation of the normal relative permeability in this range from the relative permeability of such a hypothetical linear material.

**P13.21.** Figures P13.21a and b show two hysteresis loops corresponding to sinusoidal variation of the magnetic field strength in two ferromagnetic cores between $-H_m$ and $+H_m$. Plot the time dependence of the magnetic flux density in each core. Does the magnetic flux also have a sinusoidal time dependence?



(a)

(b)
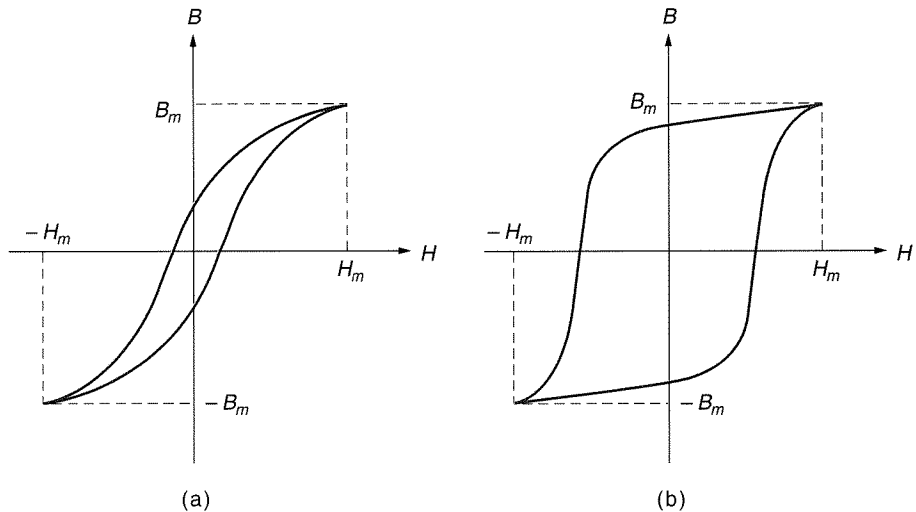
Figure P13.21 Hysteresis loops for two ferromagnetic materials

**P13.22.** If in Figs. P13.21a and b the magnetic flux density varies sinusoidally in time, sketch the time dependence of the magnetic field strength in the core. Is it also sinusoidal? If the hysteresis loops were obtained by measurements with sinusoidal magnetic field strength, is it absolutely correct to use such loops in this case?

**P13.23.** The initial magnetization curve (first part of hysteresis curve) $B(H)$ of a ferromagnetic material used for a transformer was measured and it was found that it can be approximated by a function of the form $B(H) = B_0H/(H_0 + H)$, where the coefficients are $B_0 = 1.37$ T and $H_0 = 64$ A/m. Then a thin torus with mean radius $R = 10$ cm and a cross section of $S = 1$ cm$^2$ is made out of this ferromagnetic material, and $N = 500$ turns are densely wound around it. Plot $B(H)$ and the flux through the magnetic circuit as a function of current intensity $I$ through the winding. Find the flux for (1) $I = 0.25$ A, (2) $I = 0.5$ A, (3) $I = 0.75$ A, and (4) $I = 1$ A.

**P13.24.** Assume that for the ferromagnetic material in problem P13.23 you did not have a measured hysteresis curve, but you had one data point: for $H = 1000$ A/m, $B$ was measured to be $B = 2$ T. From that, you can find an equivalent permeability and solve the circuit approximately, assuming that it is linear. Repeat the calculations from the preceding problem and calculate the error due to this approximation for the four current values given in problem P13.23.

**P13.25.** The thick toroidal core sketched in Fig. P13.25 is made out of the ferromagnetic material from problem P13.23. There are $N = 200$ turns wound around the core, and the core dimensions are $a = 3$ cm, $b = 6$ cm, and $h = 3$ cm. Find the magnetic flux through the core for $I = 0.2$ A and $I = 1$ A in two different ways: (1) using the mean radius; and (2) by dividing the core into 5 layers and finding the mean magnetic field in each of the layers.
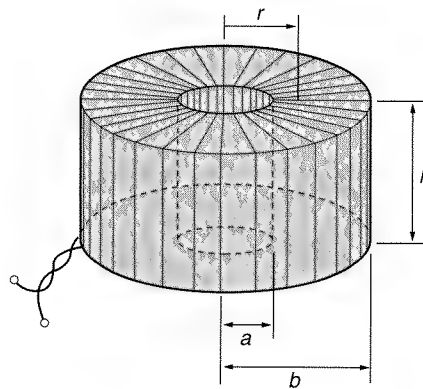


Figure P13.25 A thick toroidal coil

**P13.26.** Find the number of turns $N_1 = N_2 = N$ for the magnetic circuit shown in Fig. P13.26 so that the magnetic flux density in the air gap is $B_0 = 1$ T when $I_1 = I_2 = I = 5$ A. The core is made of the same ferromagnetic material as in problem P13.23. Solve the problem in two ways: (1) taking the magnetic resistance of the core into account; and (2) neglecting the magnetic resistance of the core. Use the following values: $a = 10$ cm, $b = 6$ cm, $d_1 = d_2 = 2$ cm, $S_1 = S_2 = S = 4$ cm$^2$, and $l_0 = 1$ mm. What is the percentage difference between the answers in (1) and (2)?
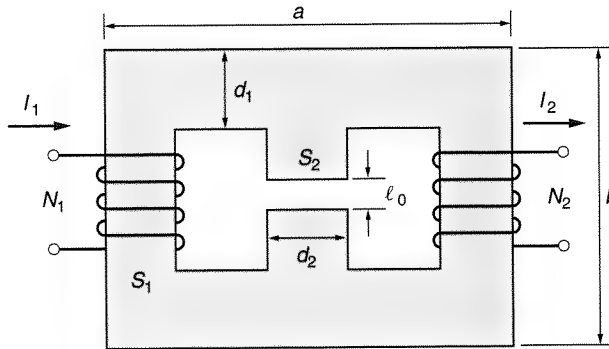
Figure P13.26 A magnetic circuit with an air gap

**P13.27.** The magnetization curve of a ferromagnetic material used for a magnetic circuit can be approximated by $B(H) = 2H/(400 + H)$, where $B$ is in T and $H$ is in A/m. The magnetic circuit has a cross-sectional area of $S = 2\,\text{cm}^2$, a mean length of $l = 50\,\text{cm}$, and $N = 200$ turns with $I = 2\,\text{A}$ flowing through them. The circuit has an air gap $l_0 = 1\,\text{mm}$ long. Find the magnetic flux density vector in the air gap.

**P13.28.** The magnetic circuit shown in Fig. P13.28 is made out of the same ferromagnetic material as in the previous problem. The dimensions of the circuit are $a = 6\,\text{cm}$, $b = 4\,\text{cm}$, $d = 1\,\text{cm}$, $S_1 = S_2 = S = 1\,\text{cm}^2$, $N_1 = 50$, $N_2 = 80$, and $N_3 = 40$. With $I_2 = I_3 = 0$, find the value of $I_1$ needed to produce a magnetic flux of $50\,\mu\text{Wb}$ in branch 3 of the circuit.
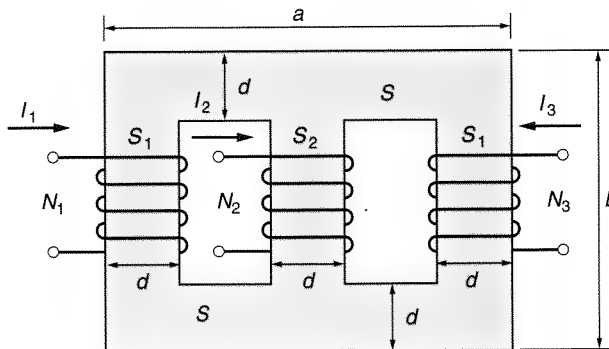


Figure P13.28 A magnetic circuit

**P13.29.** A linear magnetic circuit is shown in Fig. P13.29. The first winding has $N_1 = 100$ turns, and the second one $N_2 = 48$. Find the magnetic flux in all the branches of the circuit if the currents in the windings are (1) $I_1 = 10\,\text{mA}$, $I_2 = 10\,\text{mA}$; (2) $I_1 = 20\,\text{mA}$, $I_2 = 0\,\text{mA}$; (3) $I_1 = -10\,\text{mA}$, $I_2 = 10\,\text{mA}$. The magnetic material of the core has $\mu_r = 4000$, the dimensions of the core are $a = 4\,\text{cm}$, $b = 6\,\text{cm}$, $c = 1\,\text{cm}$, and the thickness of the core is $d = 1\,\text{cm}$.
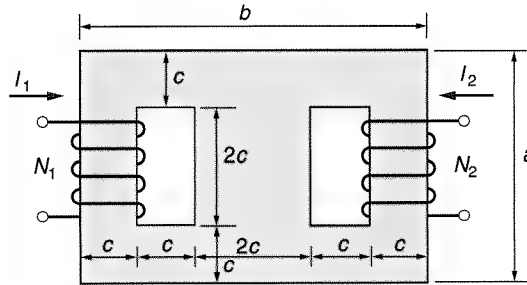
Figure P13.29 A linear magnetic circuit

**P13.30.** Shown in Fig. P13.30 is a single current loop on a toroidal core (indicated in dashed lines) of very high permeability. Assume that the core can be obtained by a gradual increase of the number of the Ampère currents, from zero to the final number per unit volume. Follow the process of creating the magnetic field in the core as the core becomes "denser." If your reasoning is correct, you should come to the answer to an important question: what is the physical mechanism of channeling the magnetic flux by ferromagnetic cores?
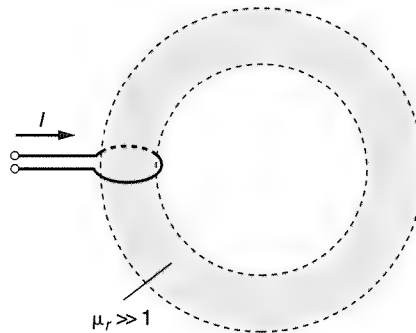


Figure P13.30 A single loop on a toroidal core

# 14

# Electromagnetic Induction and Faraday's Law

## 14.1 Introduction

We mentioned earlier that in 1831 Michael Faraday performed experiments to check whether current is produced in a closed wire placed near a magnet, reciprocally to dc currents producing magnetic fields. He found no current in that case, but realized that *a time-varying current in the loop is obtained while the magnet is being moved toward or away from it.* The law he formulated is known as *Faraday's law of electromagnetic induction.* It is perhaps the most important law of electromagnetism. Without it there would be no electricity from rotating generators, no telephone, no radio and television, to mention but a few applications.

The phenomenon of electromagnetic induction has a simple physical interpretation. We now know that two charged particles at rest act on each other with a force given by Coulomb's law. We also know that two charges moving with uniform velocities act on each other with an additional force, the *magnetic force.* If a particle is *accelerated*, it turns out that it exerts yet another force on other charged particles, stationary or moving. As in the case of the magnetic force, if only a pair of charges is considered, this additional force is much smaller than Coulomb's force. However, time-varying currents in conductors involve a vast number of accelerated charges, so they produce effects significant enough to be easily measured.

238

This additional force is of the same *form* as the electric force ($\mathbf{F} = Q\mathbf{E}$), but the electric field vector $\mathbf{E}$ in this case has quite different properties than the electric field vector of static charges. When we wish to stress this difference, we use a slightly different name—the *induced electric field strength*. This chapter introduces the concept of the induced electric field, explains the phenomenon of electromagnetic induction, and provides a derivation of Faraday's law.

## 14.2   The Induced Electric Field

To understand electromagnetic induction, we need to reconsider the concepts of electric and magnetic fields.

A dc current $I$ flowing through a stationary contour $C$ in a coordinate system $(x, y, z)$ produces a magnetic flux density field $\mathbf{B}$. Let us look at a charged particle $Q$ moving at a velocity $\mathbf{v}$ with respect to contour $C$. We add a second coordinate system $(x', y', z')$ that moves together with the charge $Q$, that is, with respect to which $Q$ is stationary.

In our thought experiment we have two observers (electrical engineers or physicists, of course), one stationary in $(x, y, z)$, and the other in $(x', y', z')$. They are interested in measuring the electric and magnetic forces acting on the charged particle, as sketched in Fig. 14.1.

Let Jack be in the first coordinate system. His instruments record a force acting on a *moving* particle. He concludes that the charge is experiencing a magnetic force $\mathbf{F} = Q\mathbf{v} \times \mathbf{B}$, since it is moving in a *time-invariant* magnetic field. If the charge stops, there is no force. Therefore, Jack's conclusion is that in his system there is no electric field.

Jill, in the second coordinate system, comes to a different conclusion. She also measures a force, proportional to $Q$, acting on the charge. However, for her the charge
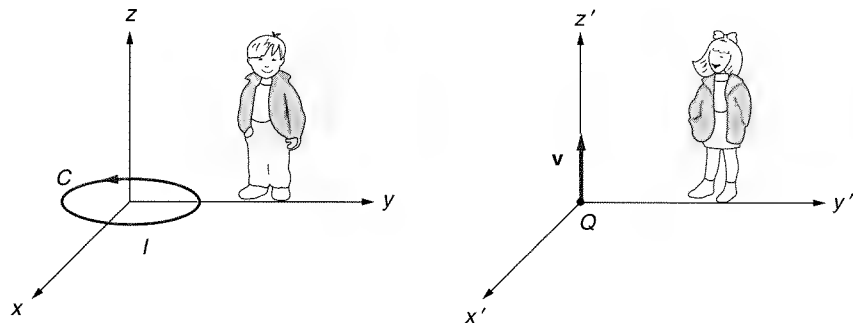


Figure 14.1 Jack and Jill are associated with different coordinate systems that are moving with respect to each other. A charged particle is at rest with respect to Jill. The two observers will explain forces measured on the charged particle in very different ways. (This picture, with the explanation given, gives a good physical insight into induced electric fields. It is worth spending some time thinking about this and understanding it.)

is *not moving*. Therefore, she concludes that the force she measured is an electric one, $\mathbf{F} = Q\mathbf{E}$. She notices, of course, that this force is time-varying. She also notices that in her system there is a time-varying magnetic field (since the source $I$ of the magnetic field is moving with respect to her coordinate system). Thus, Jill's conclusion is that in her coordinate system both a time-varying electric field and a time-varying magnetic field exist.

This strange conclusion, that two observers moving with respect to each other explain things differently, is absolutely correct. It is based, essentially, only on the definition of the electric and magnetic fields. We say that there is an electric field in a domain if a force of the form $Q\mathbf{E}$ is acting on a stationary charge (with respect to our coordinate system). If a force of the form $Q\mathbf{v} \times \mathbf{B}$ is acting on a charge moving with a velocity $\mathbf{v}$, we say that there is a magnetic field. *There are no other definitions of the electric and magnetic fields*—alternative definitions can always be reduced to these two.

Let us rephrase the important conclusion we reached: a time-varying magnetic field is accompanied by a time-varying electric field. We found this to be true in the case of motion of the observer with respect to the source of a time-invariant magnetic field. We shall now argue that a time-varying magnetic field is always accompanied by a time-varying electric field, no matter what the cause of the variation of the field is.

Assume that the source of the magnetic field is a very long and densely wound solenoid with a current $I$, as in Fig. 14.2. First Jill and the charge $Q$ do not move with respect to the solenoid. Because the charge is not moving with respect to the source of the magnetic field, and the field is constant in time, no force is acting on the charge.

Consider now the following two ways of changing the magnetic field. Let us first move the solenoid periodically between the positions 1 and 2 indicated in the figure. The charge $Q$ and Jill are now moving with respect to the solenoid. Therefore, from the viewpoint of Jack sitting on the solenoid, a magnetic force is acting on the charge. Jill, in her coordinate system, observes a changing electric force on the charge at rest with respect to her, i.e., a changing electric field and a changing magnetic field.

Suppose that instead of moving the solenoid, we move the sliding contact in the figure back and forth between positions 1 and 2. Since this turns the current on
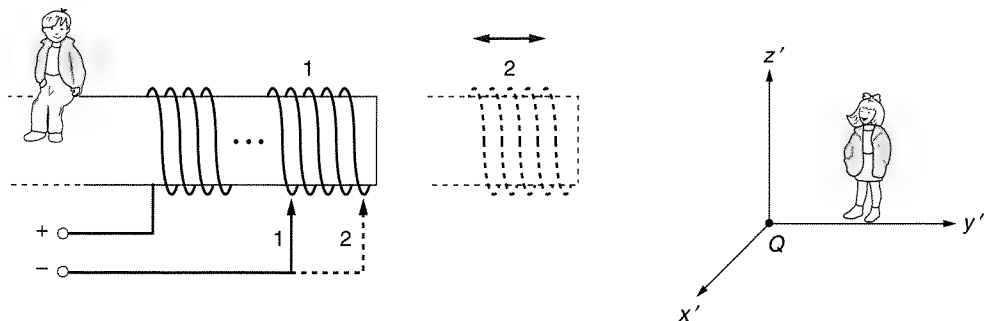


Figure 14.2 Different ways of changing the magnetic field of a solenoid lead to the same conclusion: a time-varying electric field always accompanies a time-varying magnetic field.

and off in successive solenoid windings, the magnetic field changes in the same way as before. However, the mechanism is different: we do not move the source of the magnetic field, but change the current in the source (i.e., turn it on and off in some of the windings). For Jill, who observes only the currents producing the field, there is no difference whatsoever, because the windings with the current move as before, back and forth, as far as Jill is concerned. She will find the same time-varying magnetic field as before, and the same electric force on the charge. From this simple example we infer that *no matter what the cause of the time-varying magnetic field is, a time-varying electric field is associated with it.*

This time-varying electric field is called the *induced electric field*. It is defined by the measured force on a particle $\mathbf{F} = Q\mathbf{E}$, as is the static electric field. We shall see, however, that it has quite different properties.

Of course, a charge can be situated simultaneously in both a static (Coulomb-type) and an induced field. In that case we would measure the total force

$$\mathbf{F} = Q(\mathbf{E}_{st} + \mathbf{E}_{ind}). \tag{14.1}$$

How can we determine the expression from which it is possible to evaluate the induced electric field strength? When a charged particle is moving with a velocity $\mathbf{v}$ with respect to the source of the magnetic field, the answer is simple: from the preceding example, the induced electric field is obtained as

$$\mathbf{E}_{ind} = \mathbf{v} \times \mathbf{B} \qquad (V/m). \tag{14.2}$$

*(Induced electric field observed by an observer moving with velocity $\mathbf{v}$ with respect to the source of the magnetic field)*

We concluded that time-varying currents are also sources of the induced electric field (as well as of a time-varying magnetic field). As in the case of magnetic forces, due to the small magnitude of the induced field of a single charge when compared with the Coulomb field, it is not possible to find the expression for the induced field of a single charge experimentally. However, the induced field of time-varying currents is large enough to be easily measured.

Assume we have a current distribution of density $\mathbf{J}$ (a function of time and position) in a vacuum, localized inside a volume $v$. The induced electric field is then found to be

$$\mathbf{E}_{ind} = -\frac{\partial}{\partial t}\left(\frac{\mu_0}{4\pi}\int_v \frac{\mathbf{J}\,dv}{r}\right) \qquad (V/m). \tag{14.3}$$

*(Induced electric field of slowly time-varying currents)*

In this equation, as usual, $r$ is the distance of the point where the induced field is being determined from the volume element $dv$. In the case of currents over surfaces, $\mathbf{J}\,dv$ in Eq. (14.3) should be replaced by $\mathbf{J}_s\,dS$, and in the case of a thin wire by $i\,d\mathbf{l}$.

If we know the distribution of time-varying currents, Eq. (14.3) enables the determination of the induced electric field at any point of interest. Most often it is not possible to obtain the induced electric field strength in analytical form, but it can always be evaluated numerically.
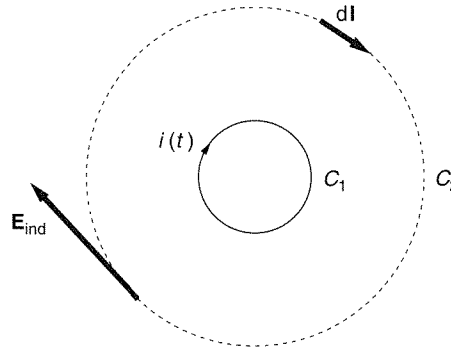
**Figure 14.3** A circular loop $C_1$ with a time-varying current $i(t)$. The induced electric field of this current is tangential to the circular loop $C_2$ indicated in dashed line, so that it results in a distributed emf around the loop.

**Example 14.1—The principle of magnetic coupling.** Let a time-varying current $i(t)$ exist in a circular loop $C_1$ of radius $a$ (Fig. 14.3). According to Eq. (14.3), lines of the induced electric field around the loop are circles, so that the line integral of the induced electric field around a circular contour $C_2$ indicated in the figure in dashed line *is not zero*. If the contour $C_2$ is a wire loop, this field will act as a distributed generator along the entire loop length, and a current will be induced in that loop.

This reasoning does not change if the loop $C_2$ is not circular. We have thus reached an extremely important conclusion: the induced electric field of time-varying currents in one wire loop produces a time-varying current in an adjacent closed wire loop. Note that the other loop need not (and usually does not) have any physical contact with the first loop. This means that the induced electric field enables transport of energy from one loop to the other *through a vacuum*. Although this coupling is actually obtained by means of the induced electric field, it is known as *magnetic coupling.*

Note that if the wire loop $C_2$ is not closed, the induced field nevertheless induces distributed generators along it. The loop behaves as an open-circuited equivalent (Thévenin) generator.

*Questions and problems:*    Q14.1 to Q14.13, P14.1 to P14.4

## 14.3   Faraday's Law

Faraday's law is an equation for the *total* electromotive force (emf) induced in a closed loop due to the induced electric field. We know that this electromotive force is distributed along the loop, but we are rarely interested in this distribution. Thus Faraday's law gives us what is of importance only from the circuit-theory point of view—the emf of the Thévenin generator equivalent to all the elemental generators acting in the loop.
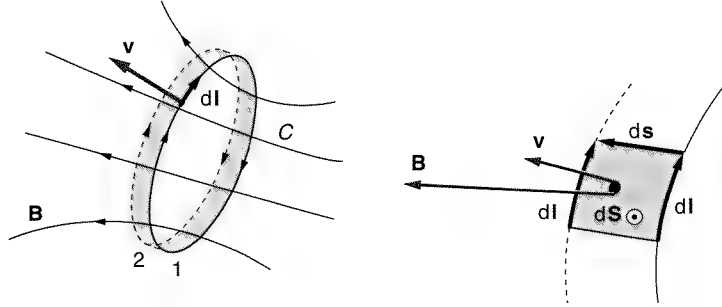
Figure 14.4 A wire loop moving in a magnetic field that is constant in time

Consider a closed conductive contour $C$ moving arbitrarily in a time-constant magnetic field (Fig. 14.4). Let us observe the contour during a short time interval $dt$. During this time, all segments $d\mathbf{l}$ of the contour move by a short distance $d\mathbf{s} = \mathbf{v}\,dt$ (different for each segment), where $\mathbf{v} = d\mathbf{s}/dt$ is the velocity of the segment considered.

Because the wire segments are moving in a magnetic field, there is an induced field acting along them of the form in Eq. (14.2). As a result, a segment behaves as an elemental generator of an emf $de = (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l}$. The emf induced in the entire contour is given by

$$e = \oint_C \mathbf{E}_{ind} \cdot d\mathbf{l} = \oint_C (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l} = \oint_C \left( \frac{d\mathbf{s}}{dt} \times \mathbf{B} \right) \cdot d\mathbf{l} = \frac{d}{dt} \oint_C (d\mathbf{s} \times \mathbf{B}) \cdot d\mathbf{l}.$$

$$(14.4)$$

The right side of this equation can be transformed as follows. From vector algebra (see Appendix 1) we know that $(d\mathbf{s} \times \mathbf{B}) \cdot d\mathbf{l} = (d\mathbf{l} \times d\mathbf{s}) \cdot \mathbf{B}$. The vector $d\mathbf{l} \times d\mathbf{s} = d\mathbf{S}$ is a vector surface element shown in Fig. 14.4. So the integral on the right-hand side in Eq. (14.4) represents the magnetic flux through the hatched strip shown in the figure. Therefore, the emf induced in the contour can be written in the form

$$e = \frac{d\Phi_{strip}}{dt} \qquad (V).$$

$$(14.5)$$

From Fig. 14.4, the flux $d\Phi_{strip}$ can also be interpreted as the difference in fluxes through the contour from position 1 to position 2, $d\Phi_{strip} = \Phi_1 - \Phi_2$. On the other hand, from the definition of the increment in flux through a contour $C$ in a time interval $dt$, $d\Phi_{through\ C\ in\ dt} = \Phi_2 - \Phi_1$. We finally have

$$e = \oint_C \mathbf{E}_{ind} \cdot d\mathbf{l} = -\frac{d\Phi_{through\ C\ in\ dt}}{dt} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{S} \qquad (V).$$

$$(14.6)$$

*(Faraday's law of electromagnetic induction)*

This is *Faraday's law of electromagnetic induction*. Recall again that the induced emf in this equation is nothing but the voltage of the Thévenin generator equivalent to all the elemental generators of electromotive forces $\mathbf{E}_{ind} \cdot \mathbf{dl}$ acting *around the loop*.

We know that an induced electric field, which is the actual cause of the emf $e$, does not depend on the mechanism by which the magnetic field changed. For example, the magnetic field variation could be due to mechanical motion in the field and/or to time-variable current sources. Equation (14.6) is valid in all those cases. Note that, except for these two ways of changing the magnetic flux in time, there are no other possibilities that result in an induced emf.

**Example 14.2—An electric generator based on electromagnetic induction.** An important example of the application of Faraday's law is electric generators. A simplified generator is sketched in Fig. 14.5. A straight piece of wire can slide along two parallel wires 1 and 2 that are at a distance $a$. At one end the wires are connected by a resistor $R$. The entire system is in a uniform magnetic field with a magnetic flux density $\mathbf{B}$ perpendicular to and into the page. A mechanical force is tagging the piece of wire with a constant velocity $\mathbf{v}$ as shown.

Let us find the emf induced in the wire due to its motion in the magnetic field. The conductor $AA'$ forms a closed loop with the rails and the resistor. The induced emf in the loop according to Faraday's law is

$$e = -\frac{d\Phi}{dt} = -\frac{\mathbf{B} \cdot d\mathbf{S}(t)}{dt} = -\frac{-B\,dS(t)}{dt} = -B\frac{av\,dt}{dt} = vBa.$$

Note that we can also get this from $e = (\mathbf{v} \times \mathbf{B}) \cdot \mathbf{a}$, where $\mathbf{a}$ is defined as in the figure.

There will be a current $I = e/R = vBa/R$ in the loop $AA'RA$ due to the induced emf. The resistor power is $P = RI^2 = v^2B^2a^2/R$.

What forces act on the piece of wire sliding along the "rail"? A current $I = vBa/R$ flows through the conductor, so a magnetic force acts on elements $\mathbf{dl}$ of the conductor $AA'$. We know that this elementary magnetic force is obtained as

$$d\mathbf{F}_m = I\,\mathbf{dl} \times \mathbf{B}.$$
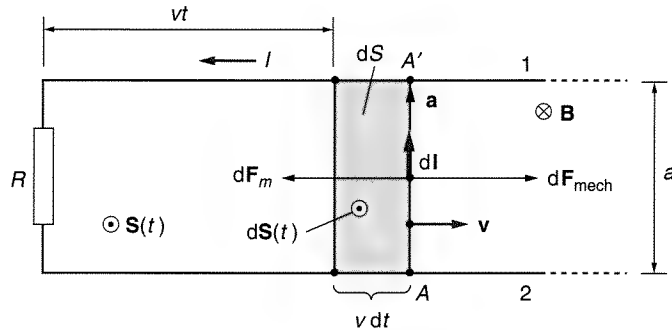


Figure 14.5 A simplified electric generator based on electromagnetic induction

Because the wire is straight and perpendicular to the magnetic flux density vector, the total magnetic force on the moving conductor is $F_m = IaB$, and its direction is as indicated in the figure. This force is *opposite* to the mechanical force that is moving the wire. Since the velocity **v** is constant, the mechanical and magnetic forces have to be equal, so $F_{mech} = IaB$. The mechanical power is $P_{mech} = F_{mech}v$, and it must be equal to the power dissipated in the resistor. Thus the velocity at which the conductor is moving is

$$v = \frac{P_{mech}}{F_{mech}} = \frac{P}{F_{mech}} = \frac{v^2 B^2 a^2 / R}{IaB} = \frac{v^2 Ba}{IR},$$

so that finally

$$v = \frac{IR}{Ba}.$$

This generator is not a practical one, but it shows in a simple way the principle of a generator based on electromagnetic induction. (The arrangement in Fig. 14.5 can be altered into a "magnetic rail gun," where the moving conductor is the bullet, if the resistor is replaced by a voltage source. Can you explain how such a gun would work?)

**Example 14.3—An ac generator.** Another example of Faraday's law is the ac generator sketched in Fig. 14.6a. A rectangular wire loop is rotating in a uniform magnetic field (for example, between the poles of a magnet). We can measure the induced voltage in the wire by connecting a voltmeter between contacts $C_1$ and $C_2$. **B** is perpendicular to the contour axis. The loop is rotating about this axis with an angular velocity $\omega$. If we assume that at $t = 0$ vector **B** is parallel to vector **n** normal to the surface of the loop (see Fig. 14.6a), the flux through the contour at that time is maximal and equal to $\Phi = Bab$. As the contour rotates, the flux becomes smaller. When **B** is parallel to the contour, the flux is zero, then it becomes negative, and so on. Since $\mathbf{B} \cdot \mathbf{n} = B \cos \omega t$, we can write

$$\Phi(t) = S\mathbf{B} \cdot \mathbf{n} = Bab \cos \omega t.$$
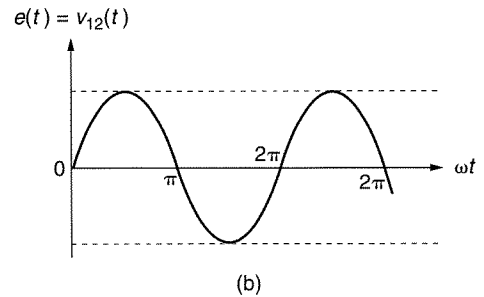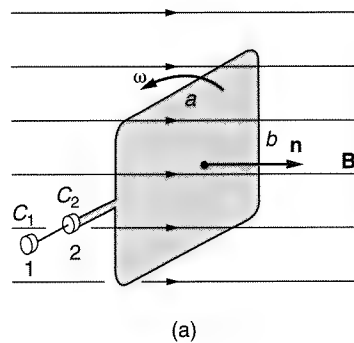


Figure 14.6 (a) A simple ac generator, and (b) the induced emf in it

The induced emf is given by

$$e(t) = -\frac{\mathrm{d}\Phi(t)}{\mathrm{d}t} = \omega abB \sin \omega t = E_{\max} \sin \omega t.$$

Its shape as a function of time is sketched in Fig. 14.6b.

A real generator has a coil with many turns of wire instead of a single loop, to obtain a larger induced emf. Also, usually the coil is not rotating; instead the magnetic field is rotating around it. This avoids sliding contacts of the generator, like those in Fig. 14.6a.

**Example 14.4—DC generator with a commutator.** The described ac generator can be modified to a dc generator. For that purpose, two sliding ring contacts are replaced by a single sliding ring contact, cut in two mutually insulated halves, as in Fig. 14.7. Such a sliding contact is known as a *commutator*. As the contour and commutator turn, the voltage between contacts 1 and 2 will always be positive because the half of the cut ring in contact with 1 will always be at a positive potential (corresponding to the parts of the sine curve in Fig. 14.6b above the $\omega t$ axis). The induced emf is shown in Fig. 14.8a. If we wanted to make this voltage less variable in time, we could have three loops at 120 degrees between successive loops. In such a case the induced emf is of the form shown in Fig. 14.8b.
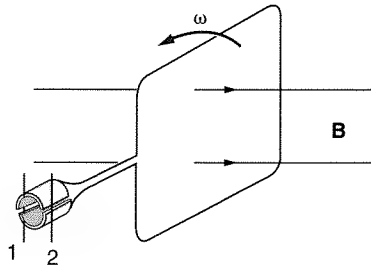


Figure 14.7 A simple dc generator with a commutator
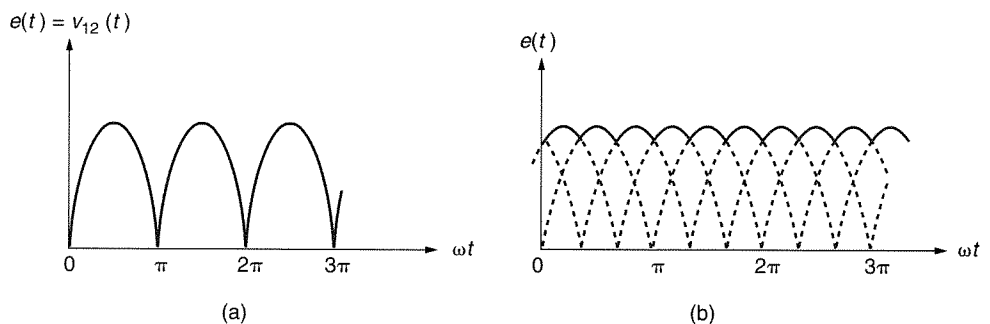


Figure 14.8 (a) The induced emf between contacts 1 and 2 of the generator from Fig. 14.7, and (b) the emf for three contours oriented at 120 degrees with respect to each other
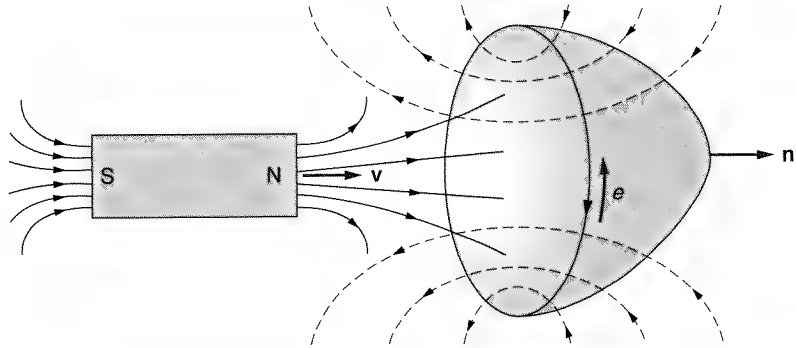
**Figure 14.9** Illustration of Lentz's law

**Example 14.5—Lentz's law.** Consider a permanent magnet approaching a stationary loop, as shown in Fig. 14.9. The permanent magnet is equivalent to a system of macroscopic currents, and, because it is moving, the magnetic flux created by these currents through the contour varies in time. According to the reference direction of the contour shown in the figure, the change of flux is positive, $d\Phi/dt > 0$, so the induced emf is in the direction shown in the figure. The emf produces a current through the closed loop, which in turn produces its own magnetic field, shown in the figure in dashed line. As a result, the change of the magnetic flux, caused initially by the magnet motion, is reduced. This is *Lentz's law*: the induced current in a conductive contour tends to decrease the change of the magnetic flux through the contour.

**Example 14.6—Eddy currents.** When a wire loop finds itself in a time-varying magnetic field, a current is induced in it due to a time-varying induced electric field that always accompanies a time-varying magnetic field. A similar thing happens in solid conductors. In a metal body we can imagine many conductive loops. A current is induced throughout the body when it is situated in a time-varying magnetic field.

The induced currents inside conductive bodies that are a result of the induced electric field are called *eddy currents*. As the first consequence of eddy currents, power is lost to heat according to Joule's law. As the second consequence, there is a secondary magnetic field due to the induced currents that reduces, by Lentz's law, the magnetic field inside the body. Both of these effects are usually not desirable. For example, in a ferromagnetic core shown in Fig. 14.10, Lentz's law tells us that eddy currents tend to decrease the flux in the core, and the magnetic circuit of the core will not be used efficiently. The flux density vector is the smallest at the center of the core, because there the **B** field of all the induced currents adds up. The total magnetic field distribution in the core is thus *nonuniform*.

To reduce these two undesirable effects, ferromagnetic cores are made of mutually insulated thin sheets, as shown in Fig. 14.11. Now the flux through the sheets is encircled by much smaller loops, the emf induced in these loops is consequently much smaller, and so the eddy currents are also reduced significantly. Of course, this only works if the vector **B** is parallel to the sheets. It is left as an exercise for the reader to explain this statement.

In some instances, eddy currents are created on purpose. For example, in so-called induction furnaces for melting metals, eddy currents are used to heat solid metal pieces to high melting temperatures.
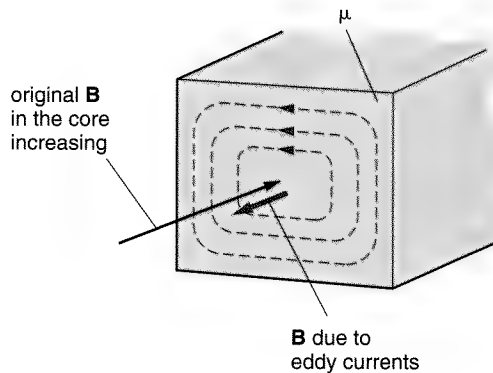
Figure 14.10 Eddy currents in a piece of ferromagnetic core. Note that the total **B** field in the core is reduced due to the opposite field created by eddy currents.

**Example 14.7—Superconducting loop.** Some substances have zero resistivity at very low temperatures. For example, lead has zero resistivity below about 7.3 K (just a little bit warmer than liquid helium). This phenomenon is known as *superconductivity,* and such conductors are said to be *superconductors.* Some ceramic materials (e.g., yttrium barium oxide) become superconductors at temperatures as "high" as about 70 K (corresponding to the temperature of liquid nitrogen). Superconducting loops have an interesting property: it is impossible to change the magnetic flux through such a loop by means of electromagnetic induction.

The explanation of this is simple. Consider a superconducting loop situated in a time-varying magnetic field. The Kirchhoff voltage law for such a loop has the form

$$-\frac{\mathrm{d}\Phi}{\mathrm{d}t} = 0,$$

since the emf in the loop is $-\mathrm{d}\Phi/\mathrm{d}t$, and the loop has zero resistance. From this equation, it is seen that the flux through a superconducting loop remains *constant.* Thus it is not possible
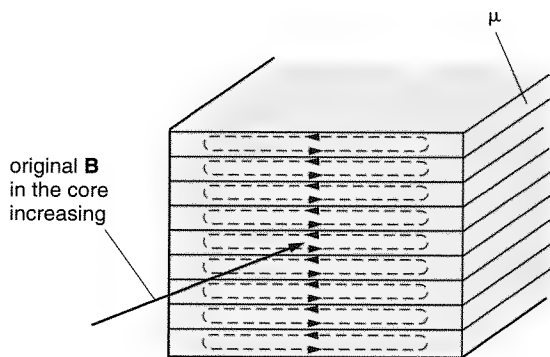


Figure 14.11 A ferromagnetic core for ac machines consists of thin insulated sheets.

to change the magnetic flux through a superconducting loop by means of electromagnetic induction.

The physical meaning of this behavior is the following: if a superconducting loop is situated in a time-varying induced electric field, the current induced in the loop must vary in time so as to produce exactly the same induced electric field in the loop, but in the opposite direction. If this were not so, infinite currents would result. We know that this induced electric field is accompanied by a time-varying magnetic field, the flux of which through the contour will be exactly the negative of the external flux.

*Questions and problems:*    Q14.14 to Q14.38, P14.5 to P14.25

## 14.4  Potential Difference and Voltage in a Time-Varying Electric and Magnetic Field

In the discussion of electrostatics, we defined the voltage to be the same as the potential difference. Actually, the voltage between two points is defined as the line integral of the *total* electric field strength from one point to the other. In electrostatics, the induced electric field does not exist, and therefore voltage is identical to potential difference. We shall now show that this is *not* the case in a time-varying electric and magnetic field.

Consider arbitrary time-varying currents and charges producing a time-varying electric and magnetic field, as in Fig. 14.12. Consider two points, *A* and *B*, in this field, and two paths, *a* and *b*, between them, as indicated in the figure. The voltage between these two points is defined as
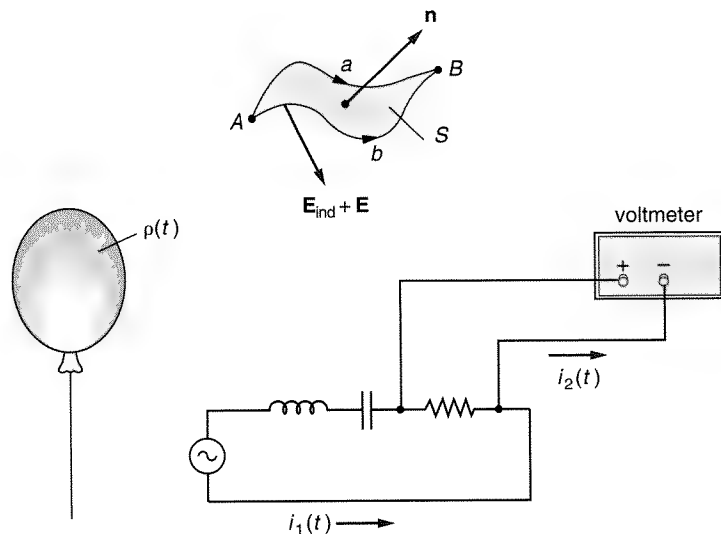


Figure 14.12 An arbitrary distribution of time-varying currents and charges

$$V_{AB} = \int_A^B \mathbf{E}_{\text{total}} \cdot d\mathbf{l}. \qquad\qquad (14.7)$$

*(Definition of voltage between two points)*

In this definition, $\mathbf{E}_{\text{total}}$ is the total electric field strength, which means the sum of the "static" part (produced by charges) and the induced part (due to time-varying currents). We know that the integral between $A$ and $B$ of the static part is simply the potential difference between $A$ and $B$. So we can write

$$V_{AB} = V_A - V_B + \int_A^B \mathbf{E}_{\text{ind}} \cdot d\mathbf{l}. \qquad\qquad (14.8)$$

We know that the potential difference, $V_A - V_B$, does not depend on the path between $A$ and $B$, but we shall now prove that the integral in this equation is different for paths *a* and *b*. These paths form a closed contour. Applying Faraday's law to that contour, we have

$$e_{\text{induced in closed contour } AaBbA} = \oint_{AaBbA} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l}$$

$$= \int_{AaB} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} - \int_{AbB} \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} = -\frac{d\Phi}{dt}, \qquad (14.9)$$

where $\Phi$ is the magnetic flux through the surface bounded by the contour $AaBbA$. Since the right side of this equation is generally nonzero, the line integrals of $\mathbf{E}_{\text{ind}}$ from $A$ to $B$ along *a* and along *b* are different. Consequently, *the voltage between two points in a time-varying electric and magnetic field depends on the particular path between these two points.*

This is an important practical conclusion. We always measure the voltage by a voltmeter with leads connected to the two points between which the voltage is being measured. Circuit theory postulates that this voltage does not depend on the shape of the voltmeter leads. We now know that in the time-varying case this is not true. Because the difference in voltage for two paths depends on the rate of change of the flux, this effect is particularly pronounced at high frequencies.

*Questions and problems:*   Q14.39, P14.26 to P14.28

## 14.5  Chapter Summary

1. If a closed wire loop is moving with respect to a source of a time-invariant magnetic field, or is near another loop with time-varying electric current, a distributed emf is induced along it. This phenomenon is known as *electromagnetic induction*. It is due to a component of the electric field, the *induced electric field*, which exists along the loop.

2. The induced electric field acts on a point charge $Q$ with a force $\mathbf{F} = Q\mathbf{E}$, like the electric field due to stationary charges, but its line integral around a closed contour is not zero. Precisely this gives rise to electromagnetic induction.

3. Integrally, i.e., not taking care of the exact distribution of the induced emf (which is of interest only rarely), the induced emf equals the negative rate of time variation of the magnetic flux through the contour. This is known as Faraday's law of electromagnetic induction.

4. The voltage between two points is defined as a line integral of the *total* electric field along a line joining them. Because of the induced electric field, the voltage depends on the particular line joining the two points, although the potential difference (which is a part of this voltage) does not.

## QUESTIONS

**Q14.1.** An observer in a coordinate system with the source of a time-invariant magnetic field observes a force $\mathbf{F} = Q\mathbf{v} \times \mathbf{B}$ on a charge $Q$ moving with a uniform velocity $\mathbf{v}$ with respect to his coordinate system. What is the force on $Q$ observed by an observer moving with the charge? What is his interpretation of the velocity $\mathbf{v}$?

**Q14.2.** Three point charges are stationary in a coordinate system of the first observer. A second observer, in his coordinate system, moves with respect to the first with a uniform velocity. What kinds of fields are observed by the first observer, and what by the second?

**Q14.3.** A small uncharged conducting sphere is moving in the field of a permanent magnet. Are there induced charges on the sphere surface? If they exist, how do an observer moving with the charge and an observer on the magnet explain their existence?

**Q14.4.** A straight metal rod moves with a constant velocity $\mathbf{v}$ in a uniform magnetic field of magnetic induction $\mathbf{B}$. The rod is normal to $\mathbf{B}$, and $\mathbf{v}$ is normal to both the rod and to $\mathbf{B}$. Sketch the distribution of the induced charges on the rod. What is the *electric* field of these charges inside the rod equal to?

**Q14.5.** A small dielectric sphere moves in a uniform magnetic field. Is the sphere polarized? Explain.

**Q14.6.** A charge $Q$ is located close to a toroidal coil with time-varying current. Is there a force on the charge? Explain.

**Q14.7.** A wire of length $l$ is situated in a magnetic field of flux density $\mathbf{B}$ parallel to the wire. Is an electromotive force induced in the wire if it is moved (1) along the lines of $\mathbf{B}$, and (2) transverse to the lines of $\mathbf{B}$? Explain.

**Q14.8.** Strictly speaking, do currents in branches of an ac electric circuit depend on the circuit shape? Explain.

**Q14.9.** Does the shape of a dc circuit influence the currents in its branches? Explain.

**Q14.10.** A circular metal ring carries a time-varying current, which produces a time-varying induced electric field. If the ring is set in oscillatory motion about the axis normal to its plane, will the induced electric field be changed? Explain your answer.

**Q14.11.** A device for accelerating electrons known as the *betatron* (Fig. Q14.11) consists of a powerful electromagnet and an evacuated tube that is bent into a circle. Electrons are accelerated in the tube when the magnetic flux in the core is forced to rise approxi-
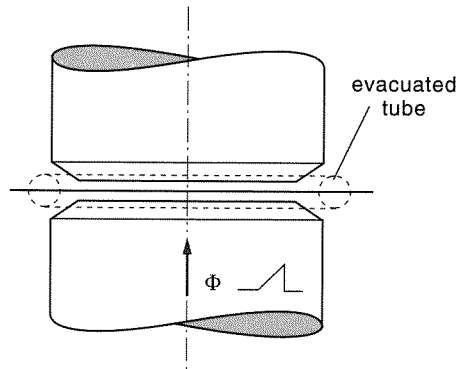
Figure Q14.11 Sketch of a betatron

mately as a linear function of time. What is the physical mechanism for accelerating the electrons in the betatron?

**Q14.12.** Is it physically sound to speak about a partial electromotive force induced in a segment of one loop by the current in a segment of another (or even of the same) loop? Explain.

**Q14.13.** A vertical conducting sheet (say, of aluminum) is permitted to fall under the action of gravity between the poles of a powerful permanent magnet. Is the motion of the sheet affected by the presence of the magnet? Explain.

**Q14.14.** A long solenoid wound on a Styrofoam core carries a time-varying current. It is encircled by three loops, one of copper, one of a resistive alloy, and the third of a bent moist filament (a poor conductor). In which loop is the induced electromotive force the greatest?

**Q14.15.** What becomes different in question Q14.14 if the solenoid is wound onto a ferromagnetic core?

**Q14.16.** Assume that the current $i(t)$ in circular loop 1 in Fig. Q14.16 in a certain time interval increases linearly in time. Will there be a current in the closed conducting loop 2? If you think that there will be, what is its direction?
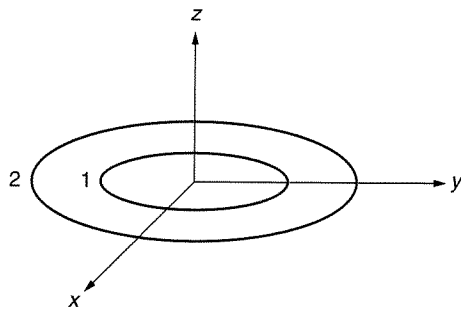


Figure Q14.16 Two coupled coils

**Q14.17.** What is the direction of the current induced in the loop sketched in Fig. Q14.17? Explain.
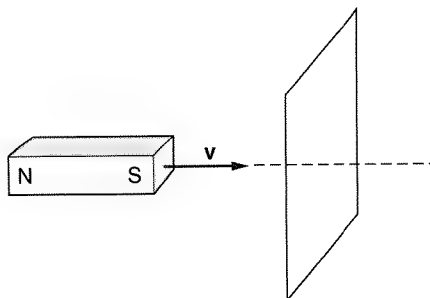
Figure Q14.17 A permanent magnet approaching a loop
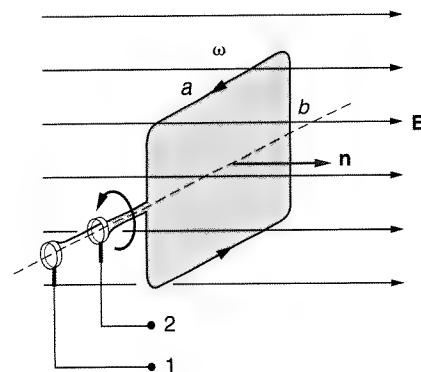
Figure Q14.18 A coil rotating in a magnetic field

**Q14.18.** The coil in Fig. Q14.18 consists of $N$ densely wound turns of thin wire. What is the voltage of the generator when compared with the case of a single turn? Explain using both the concept of magnetic flux and that of the induced electric field.

**Q14.19.** A solenoid is wound onto a long, cylindrical permanent magnet. A voltmeter is connected to one end of the solenoid, and to a sliding contact that moves and makes a contact with a larger or smaller number of turns of the solenoid. Thus the magnetic flux in the closed loop of the voltmeter will vary in time. Does the voltmeter detect a time-varying voltage (assuming that it is sensitive enough to do so)? Explain.

**Q14.20.** What is the direction of the reference unit vectors normal to the three loops in Fig. Q14.20?
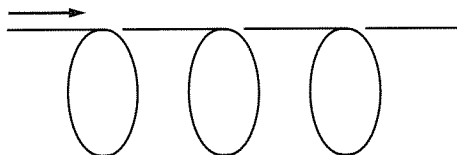
Figure Q14.20 Three series loops

**Q14.21.** A cylindrical permanent magnet falls without friction through a vertical evacuated metal tube. Is the fall accelerated? If not, what determines the velocity of the magnet? Explain.

**Q14.22.** A strong permanent magnet is brought near the end plate of the metal pendulum of a wall timepiece. Does this influence the period of the pendulum? If it does, is the pendulum accelerated or slowed down? Does this depend on the type of the magnetic pole of the magnet closer to the pendulum plate? Explain.

**Q14.23.** Figure Q14.23 shows a sketch of a flat strip moving between the poles of a permanent magnet. Sketch the lines of the induced current in the strip.
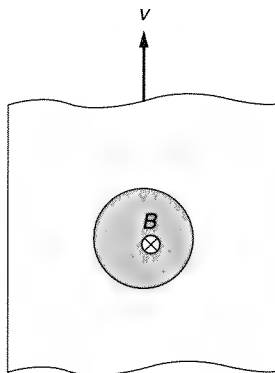


**Figure Q14.23** A strip moving in a magnetic field

**Q14.24.** The strip in question Q14.23 has in case (1) longitudinal and in case (2) transverse slots with respect to the direction of motion. In which case are induced currents greater? Explain.

**Q14.25.** Explain in detail how the right-hand side in Eq. (14.4) is obtained from the middle expression in it.

**Q14.26.** The magnetic flux through a contour $C$ at time $t$ is $\Phi_1$, and at time $t + \Delta t$ it is $\Phi_2$. Is the time increment of the flux through $C$ ($\Phi_2 - \Phi_1$), or ($\Phi_1 - \Phi_2$)? Explain.

**Q14.27.** Is the *distribution* of the induced electromotive force around a contour seen from the right-hand side in Eq. (14.6)? Is it seen from the middle expression in that equation? Is it seen from any expression in Eq. (14.4)?

**Q14.28.** Imagine an electric circuit with several loops situated in a slowly time-varying magnetic (and induced electric) field. Can you analyze such a circuit by circuit-theory methods? If you think you can, explain in detail how you would do it.

**Q14.29.** Does it make any sense at all to speak about the electromotive force induced in an *open* loop? If you think that this makes sense, explain what happens.

**Q14.30.** Explain in detail what a positive and what a negative electromotive force in Eq. (14.6) mean.

**Q14.31.** Why is the reduction of eddy current losses possible only if the vector **B** is parallel to a thin ferromagnetic sheet?

**Q14.32.** What is the induced electromotive force in the loop shown in Fig. Q14.32, if the magnetic field is time-varying? If the right half of the loop is turned about the $x$ axis by 180 degrees, what is then the induced electromotive force? Explain both in terms of the magnetic flux and of the induced electric field.
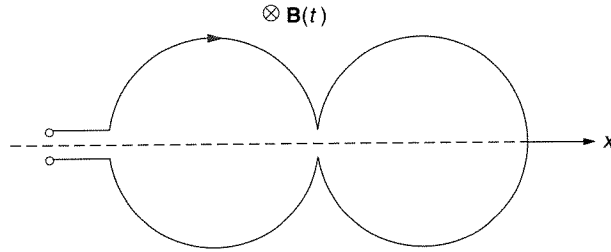
Figure Q14.32 A loop in the form of an 8

**Q14.33.** A solid conducting body is placed near a loop with time-varying current. Are any forces acting on free charges inside the body? Explain.

**Q14.34.** Is there a magnetic force between the body and the loop from the preceding example? Explain.

**Q14.35.** A planar insulated loop with time-varying current is placed on the surface of a plane conducting sheet. What happens in the sheet? Is the power required to drive the current in the loop different when it resides on the sheet than when it is isolated in space? Explain.

**Q14.36.** Of two closed conducting loops, $C_1$ and $C_2$, $C_1$ is connected to a generator of time-varying voltage. Is there a current in $C_2$? Explain.

**Q14.37.** An elastic metal circular ring carrying a steady current $I$ is periodically deformed to a flat ellipse, and then released to retain its original circular shape. Is an electromotive force induced in the loop? Explain.

**Q14.38.** In Example 14.7 it was demonstrated that the flux through a superconducting loop cannot be changed. Does this mean that the flux through a superconducting loop cannot be changed by *any* means?

**Q14.39.** Why does the shape of voltmeter leads influence the time-varying voltage the voltmeter measures? Why does this influence increase with frequency?

# PROBLEMS

**P14.1.** Starting from Eq. (14.3), prove that the lines of the induced electric field vector of a circular current loop with a time-varying current are circles centered at the loop axis.

**P14.2.** Assume that you know the induced electric field $E_{ind}(t)$ along a circular line $C$ of radius $a$ in problem P14.1. A wire loop of radius $a$ coincides with $C$. Evaluate the total electromotive force induced in the loop. Prove that this is, actually, the voltage of the Thévenin generator equivalent to the distributed infinitesimal generators around the loop.

**P14.3.** Two coaxial solenoids shown in Fig. P14.3 are connected in series. A current $i(t) = 1.5 \sin 1000t$ A, where time is in seconds, flows through the solenoids. The dimensions are $a = 1$ cm, $b = 2$ cm, and $L = 50$ cm. The number of turns in both is the same, $N_1 = N_2 = 1000$. Find the approximate induced electric field at points $A_1$ (surface of the inner solenoid), $A_2$ (halfway between the two solenoids), and $A_3$ (right outside the outer solenoid).
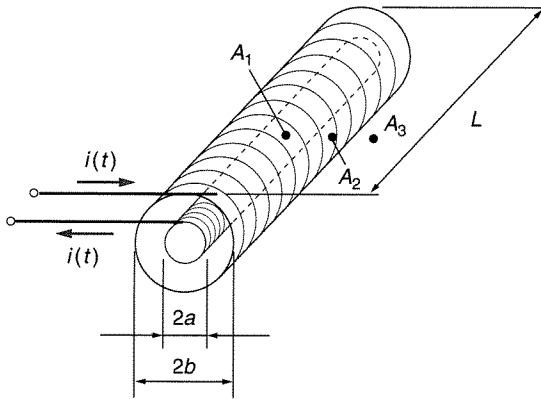
Figure P14.3 Two coaxial solenoids

**P14.4.** A circular loop of radius $a$, with a current $I$, rotates about the axis normal to its plane with an angular frequency $\omega$. A small charge $Q$ is fastened to a loop radius and rotates with the loop, as in Fig. P14.4. Is there a force on the charge? If it exists, determine its direction.
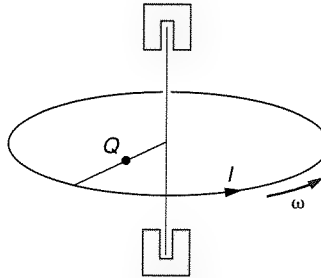


Figure P14.4 Rotating current loop and charge

**P14.5.** A side of a rectangular wire loop is partly shielded from the magnetic field normal to the loop plane with a hollow ferromagnetic cylinder, as in Fig. P14.5. Therefore the sides $ad$ and $bc$ are situated in different magnetic fields. If the loop, together with the cylinder, moves in the indicated direction with a velocity $\mathbf{v}$, will there be a current in
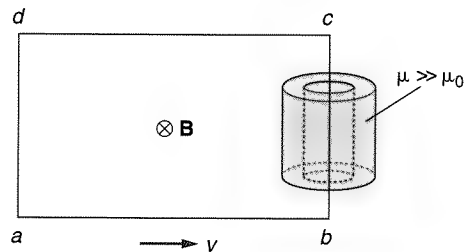


Figure P14.5 A partly shielded wire loop

the loop? If the answer is yes, could this serve for measuring the velocity with respect to the earth's magnetic field? Explain.

**P14.6.** A current $i(t) = I_m \sin(2\pi ft) = 2.5 \sin 314t$ A is flowing through the solenoid in Fig. P14.6, where frequency is in hertz and time is in seconds. The solenoid has $N_1 = 50$ turns of wire, and the coil $K$ shown in the figure has $N_2 = 3$ turns. Calculate the emf induced in the coil, as well as the amplitude of the induced electric field along the coil turns. The dimensions indicated in the figure are $a = 0.5$ cm, $b = 1$ cm, and $L = 10$ cm. Plot the induced emf as a function of $N_1$, $N_2$, and $f$.
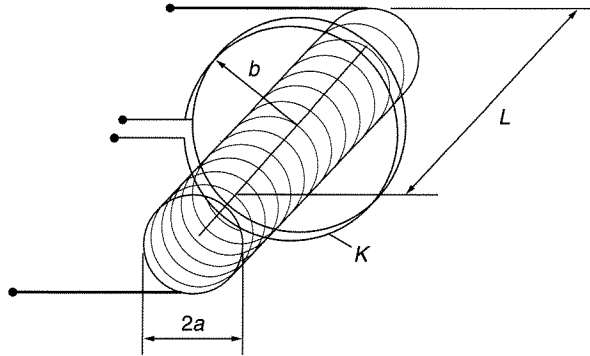


Figure P14.6 A solenoid and a coil

**P14.7.** If the conductor $AA'$ in Fig. P14.7 is rotating at a constant angular velocity and makes $n$ turns per second, find the voltage $V_{AA'}$ as a function of time. Assume that at $t = 0$ the conductor is in the position shown in the figure.
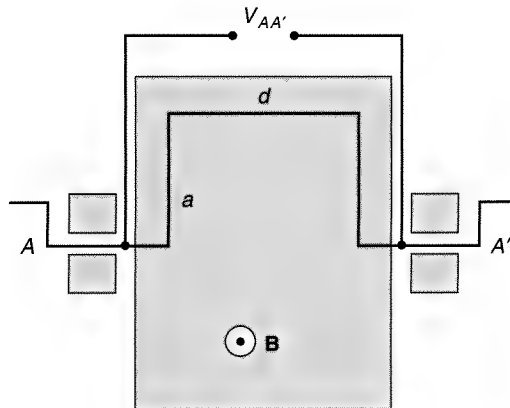


Figure P14.7 A rotating conductor

**P14.8.** A two-wire line is parallel to a long straight conductor with a dc current $I$ (Fig. P14.8). The two-wire line is open at both ends, and a conductive bar is sliding along it with a
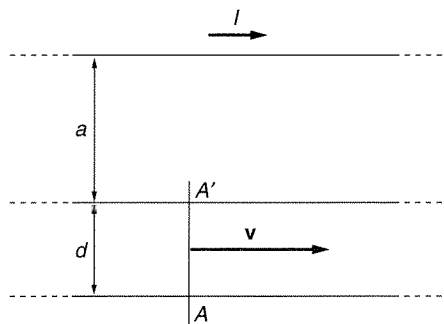
Figure P14.8 A conductor and two-wire line

uniform velocity $v$, as shown in the figure. Find the potential difference between the two line conductors.

P14.9. A rectangular wire loop with sides of lengths $a$ and $b$ is moving away from a straight wire with a current $I$ (Fig. P14.9). The velocity of the loop, $v$, is constant. Find the induced emf in the loop. The reference direction of the loop is shown in the figure. Assume that at $t = 0$ the position of the loop is defined by $x = a$.
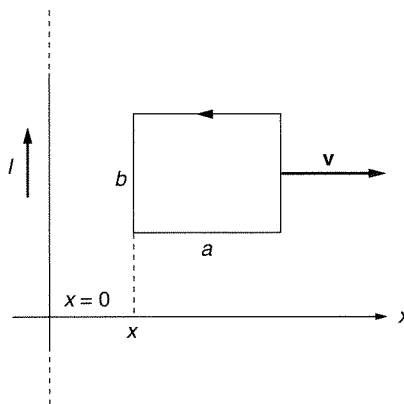


Figure P14.9 A moving frame in a magnetic field

P14.10. The current flowing through the straight wire from the preceding problem is now $i(t)$ (a function of time). Find the induced emf in the loop, which is moving away from the wire as in the preceding problem. What happens to your expression for the emf when (1) the frame stops moving, or (2) when $i(t)$ becomes a dc current, $I$?

P14.11. A liquid with a small but finite conductivity is flowing through a flat insulating pipe with an unknown velocity $v$. The velocity of the fluid is roughly uniform over the cross section of the pipe. To measure the fluid velocity, the pipe is in a magnetic field with a flux density vector **B** normal to the pipe, as shown in Fig. P14.11. Two small electrodes are in contact with the fluid at the two ends of the pipe. A voltmeter with

large input impedance shows a voltage $V$ when connected to the electrodes. Find the velocity of the fluid.
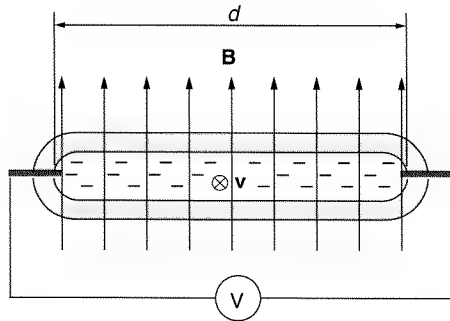


Figure P14.11 Measurement of fluid velocity
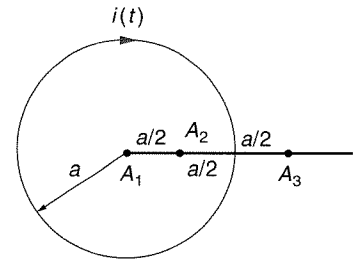


Figure P14.12 Cross section of a solenoid

**P14.12.** Shown in Fig. P14.12 is the cross section of a very long solenoid of radius $a = 1\,\text{cm}$, with $N' = 2000\,\text{turns/m}$. In the time interval $0 \le t \le 1\,\text{s}$, a current $i(t) = 50t\,\text{A}$ flows through the solenoid. Determine the acceleration of an electron at points $A_1$, $A_2$, and $A_3$ indicated in the figure. (Note: the acceleration, $\mathbf{a}$, is found from the relation $\mathbf{F} = m\mathbf{a}$, where $\mathbf{F}$ is the force on the electron.)

**P14.13.** Determine approximately the induced electric field strength inside the tubular coil sketched in Fig. P14.13. The current intensity in the coil is $I = 0.02\cos 10^6 t\,\text{A}$, the number of turns is $N = 100$, and the coil dimensions are $a = 1\,\text{cm}$, $b = 1.5\,\text{cm}$, and $L = 10\,\text{cm}$.
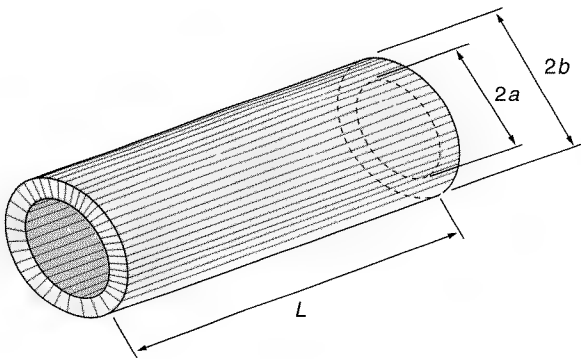


Figure P14.13 A tubular coil
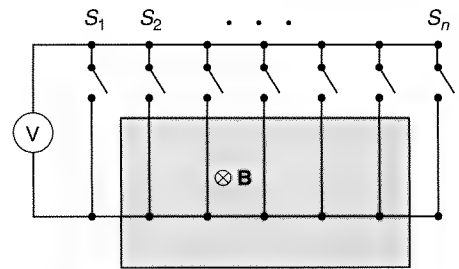


Figure P14.14 A test of electromagnetic induction

**P14.14.** Sketched in Fig. P14.14 is an experimental setup for the analysis of electromagnetic induction. By closing sequentially the switches $S_1, \ldots, S_n$, it is possible to change the magnetic flux through the closed contour shown from zero to a maximal value. Will the voltmeter indicate an emf induced in the circuit? Explain.

**P14.15.** Find the angular velocity of the rotor of an idealized electric motor shown in Fig. P14.15 for the case when no load is connected to it. Does the value of $R$ influence the angular velocity? The rotor is in the form of a metal wheel with four spokes, situated in a uniform magnetic field of magnetic flux density $\mathbf{B}$, as shown in the figure. What is the direction of rotation of the rotor?
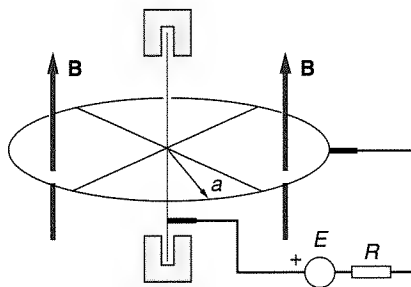
Figure P14.15 An idealized electric motor

**P14.16.** A circular loop of radius $a$ rotates with an angular velocity $\omega$ about the axis lying in its plane and containing the center of the loop. It is situated in a uniform magnetic field of flux density $\mathbf{B}(t)$ normal to the axis of rotation. Determine the induced emf in the loop. At $t = 0$, the position of the loop is such that $\mathbf{B}$ is normal to its plane.

**P14.17.** A winding of $N = 1000$ turns of wire with sinusoidal current of amplitude $I_m = 200\,\text{mA}$ is wound on a thin toroidal ferromagnetic core of mean radius $a = 10\,\text{cm}$. Figure P14.17 shows the idealized hysteresis loop of the core corresponding to the sinusoidal magnetization of the core to saturation in both directions. Also wound on the toroid are several turns of wire over the first winding. Plot the emf induced in the second winding during one period of the sinusoidal current in the first winding.
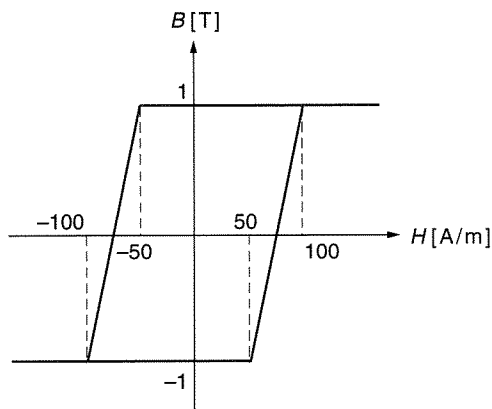
Figure P14.17 An idealized hysteresis loop

**P14.18.** Shown in Fig. P14.18 is a rectangular loop encircling a very long solenoid of radius $R$ and with $N'$ turns of wire per unit length. The amplitude of current in the winding is

$I_m$, and its angular frequency is $\omega$. Determine the emf induced in the entire rectangular loop, as well as in its sides $a$ and $b$ separately.
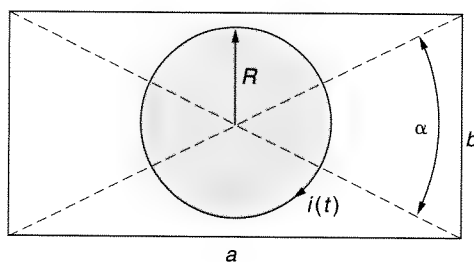


Figure P14.18 A rectangular loop encircling a solenoid

**P14.19.** The cross section of a thick coil with a large number, $N$, of turns of thin wire, is shown in Fig. P14.19. The coil is situated in a time-varying magnetic field of flux density $\mathbf{B}(t)$, in the indicated direction. Determine the emf induced in the coil.
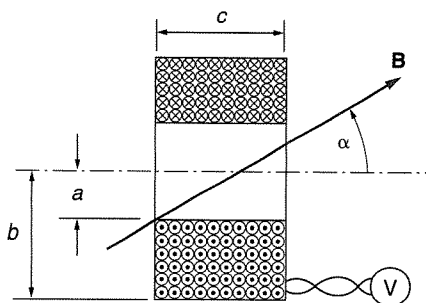


Figure P14.19 A coil of rectangular cross section
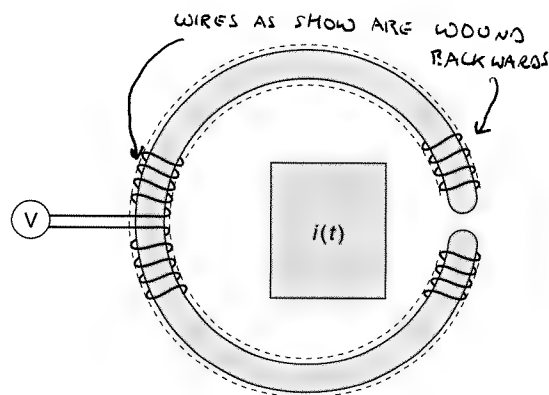


Figure P14.20 A conductor encircled by a coil

**P14.20.** The conductor whose cross section is shown shaded in Fig. P14.20 carries a sinusoidal current of amplitude $I_m$ and angular frequency $\omega$. The conductor is encircled by a flexible thin rubber strip of cross-sectional area $S$, densely wound along its length with $N'$ turns of wire per unit length. The measured amplitude of the voltage between the terminals of the strip winding is $V_m$. Determine $I_m$.

**P14.21.** Wire is being wound from a drum $D_1$ onto a drum $D$, at a rate of $N'$ turns per unit time (Fig. P14.21). The end of the wire on drum $D$ is fastened to the ring $R$, which has a sliding contact $F$. A voltmeter is connected between the contact $F$ and another contact $G$. Through the drum $D$ there is a constant flux $\Phi$, as indicated. What is the electromotive force measured by the voltmeter?
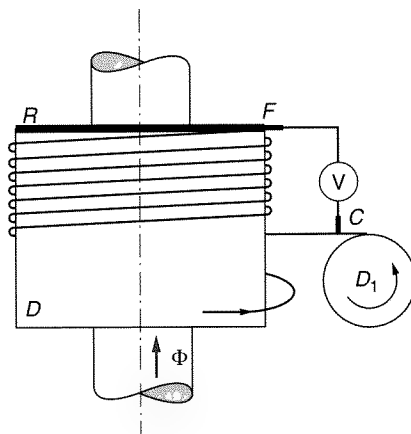
Figure P14.21 A test of electromagnetic induction

**P14.22.** A cylindrical coil is tightly wound around a ferromagnetic core with time-varying magnetic flux $\Phi(t) = \Phi_m \cos \omega t$, as shown in Fig. P14.22. The length of the coil is $L$, and the number of turns in the coil is $N$. If a sliding contact $K$ moves along the coil according to the law $x = L(1 + \cos \omega_1 t)/2$, what is the time dependence of the electromotive force between contacts $A$ and $K$? Plot your result.
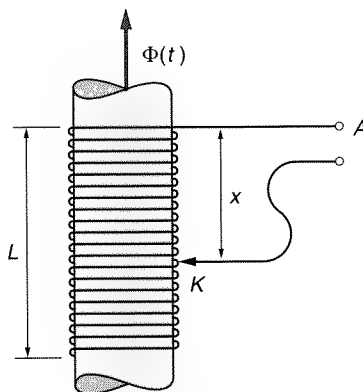


Figure P14.22 A coil with a sliding contact

**P14.23.** A cylindrical conducting magnet of circular cross section rotates about its axis with a uniform angular velocity. A galvanometer $G$ is connected to the equator of the magnet and to the center of one of its bases by means of sliding contacts, as shown in Fig. P14.23. If such an experiment is performed, the galvanometer indicates a certain current through the circuit (Faraday, 1832). Where is the electromotive force induced: in the stationary conductors connecting the sliding contacts with the galvanometer, or in the magnet itself?
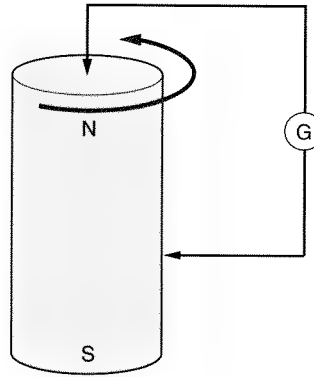
Figure P14.23 A rotating magnet with sliding contacts

**P14.24.** A ferromagnetic toroid with no air gap is magnetized so that no magnetic field exists outside it. The toroid is encircled by an elastic metal loop, as in Fig. P14.24a. The loop is now taken from the toroid in such a way that during the process, the loop is always electrically closed through the conducting material of the toroid, as in Figs. P14.24b and c. The magnetic flux through the contour was obviously changed from a value $\Phi$, the flux through the toroid, to zero. A formal application of Faraday's law leads to the conclusion that a certain charge will flow through the circuit during the process, but in this case it is not possible to detect any current (Herring, 1908). Explain the negative result of the experiment.



(a)               (b)               (c)
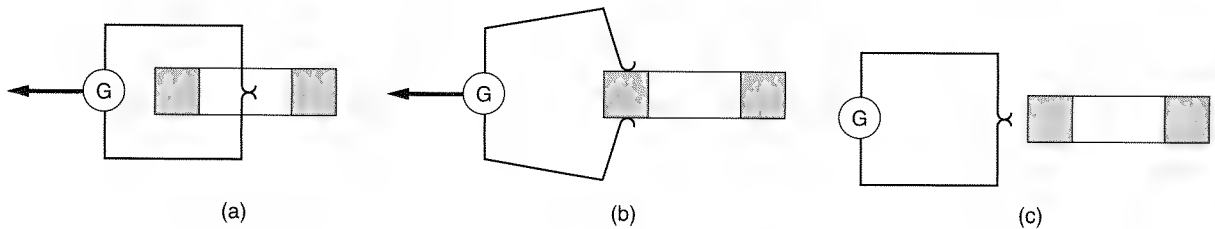
Figure P14.24 (a) A moving elastic loop encircles a magnetized toroid. (b) The loop is moving, but is closed by the conducting toroid body. (c) The loop does not move and does not encircle the toroid.

**P14.25.** Discuss the possibility of constructing a generator of electromotive force constant in time, operating on the basis of electromagnetic induction.

**P14.26.** In a straight copper wire of radius $a = 1\,\text{mm}$ there is a sinusoidal current $i(t) = 1\cos\omega t$ A. A voltmeter is connected between points 1 and 2, with leads of the shape shown in Fig. P14.26. If $b = 50\,\text{cm}$ and $c = 20\,\text{cm}$, evaluate the voltage measured by the voltmeter for (1) $\omega = 314\,\text{rad/s}$, (2) $\omega = 10^4\,\text{rad/s}$, and (3) $\omega = 10^6\,\text{rad/s}$. Assume that the resistance of the copper conductor per unit length, $R'$, is approximately that for a dc current (which actually is *not* the case, due to the so-called skin effect), and evaluate for the three cases the difference between the potential difference $V_1 - V_2 = R'bi(t)$ and the voltage induced in the leads of the voltmeter.
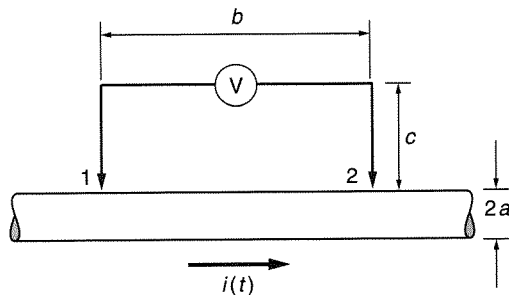
Figure P14.26 Measurement of ac voltage

**P14.27.** A circular metal loop of radius $R$, conductivity $\sigma$, and cross-sectional area $S$ encircles a long solenoid with a time-varying current $i(t)$ (Fig. P14.27). The solenoid has $N'$ turns of wire per unit length, and its radius is $r$. Determine the current in the loop, and the voltage between points $A$ and $B$ of the loop along paths $a$, $b$, $c$, and $d$. Neglect the induced electric field in the loop due to the loop current itself. Determine also the voltage between points $A$ and $C$, along the path $AcC$, and along the path $AbBC$.
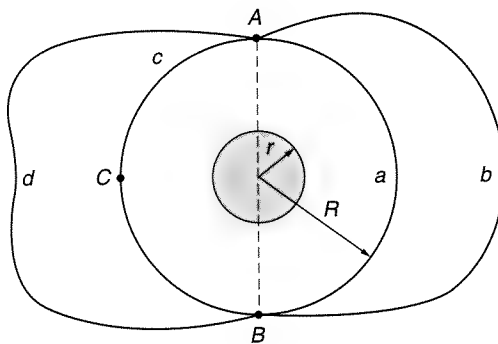


Figure P14.27 A solenoid encircled by a circular loop

**P14.28.** Repeat problem P14.27 assuming that the two halves of the loop have different conductivites, $\sigma_1$ and $\sigma_2$, and they meet at points $A$ and $B$.

# 15

# Inductance

## 15.1 Introduction

We now know that a time-varying current in one wire loop induces an emf in another loop. We do not know, however, how to compute that emf. In linear media, an electromagnetic parameter that enables simple determination of this emf is the *mutual inductance*.

A wire loop with time-varying current creates a time-varying induced electric field not only in the space around it but also along the loop itself. As a consequence, we have a kind of feedback—the current produces an effect that affects itself. The parameter known as inductance, or *self-inductance*, of the loop enables simple evaluation of this effect.

Mutual inductance and self-inductance are familiar because they are used in circuit theory for describing magnetic coupling. Many manifestations of magnetic coupling are not as familiar, however, although they are not unimportant. For example, what we call magnetic coupling can exist between a 60-Hz power line and a human body, or between parallel printed strips of a computer bus. With a knowledge of the induced electric field, these and related phenomena can be easily understood.

## 15.2 Mutual Inductance

Consider two stationary thin conductive contours $C_1$ and $C_2$ in a linear medium (e.g., air), shown in Fig. 15.1. When a time-varying current $i_1(t)$ flows through the first
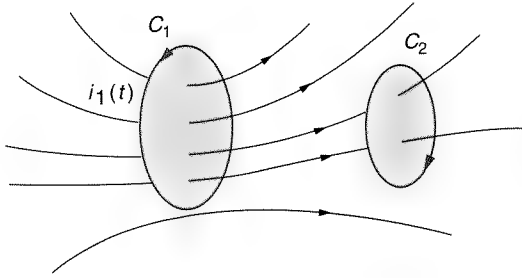
Figure 15.1 Two coupled conductive contours

contour, it creates a time-varying magnetic field as well as a time-varying induced electric field, $\mathbf{E}_{1\text{ind}}$. The latter produces an emf $e_{12}(t)$ in the second contour, given by

$$e_{12}(t) = \oint_{C_2} \mathbf{E}_{1\text{ind}} \cdot d\mathbf{l}_2, \tag{15.1}$$

*(Calculation of distributed emf along a loop)*

where the first index denotes the source of the field (contour 1 in this case).

As mentioned earlier, it is usually much easier to find the induced emf using Faraday's law than in any other way. The magnetic flux density vector in linear media is proportional to the current that causes the magnetic field. It follows that the flux $\Phi_{12}(t)$ through $C_2$ caused by the current $i_1(t)$ in $C_1$ is also proportional to $i_1(t)$:

$$\Phi_{12}(t) = L_{12}i_1(t). \tag{15.2}$$

The proportionality constant, $L_{12}$, is called the *mutual inductance* between the two contours. This constant depends only on the geometry of the system and the properties of the (linear) medium surrounding the current contours. Mutual inductance is denoted by both $L_{12}$ (or whatever subscripts are chosen) and—particularly in circuit theory—by $M$.

Because the variation of $i_1(t)$ can be arbitrary, the same expression holds when the current through $C_1$ is a dc current:

$$\Phi_{12} = L_{12}I_1. \tag{15.3}$$

*(Flux definition of mutual inductance)*

Although mutual inductance has no practical meaning for the case of dc currents, this definition is frequently used for the determination of mutual inductance.

According to Faraday's law, the emf can alternatively be written as

$$e_{12}(t) = -\frac{d\Phi_{12}(t)}{dt} = -L_{12}\frac{di_1(t)}{dt}. \tag{15.4}$$

*(EMF definition of mutual inductance)*

The unit for inductance, equal to a Wb/A, is called a henry (H). (Joseph Henry was an American physicist who independently discovered electromagnetic induction at almost the same time as Faraday did.) One henry is quite a large unit. Most frequent values of mutual inductance are on the order of a mH, $\mu$H, or even nH.

If we now assume that a current $i_2(t)$ in $C_2$ causes an induced emf in $C_1$, we talk about a mutual inductance $L_{21}$. It turns out that $L_{12} = L_{21}$ always. [This follows from the expression for the induced electric field in Eq. (15.3) and Eqs. (15.1) and (15.4), but the proof will not be given here.] So we can write[14]

$$L_{12} = \frac{\Phi_{12}}{I_1} = L_{21} = \frac{\Phi_{21}}{I_2} \qquad \text{(H)}. \tag{15.5}$$

These equations show that we need to calculate either $\Phi_{12}$ or $\Phi_{21}$ to determine the mutual inductance. In some instances one of these is much simpler to calculate, as the following example shows.

**Example 15.1—Mutual inductance of a toroidal coil and a wire loop.** Let us find the mutual inductance between a contour $C_1$ and a toroidal coil $C_2$ with $N$ turns, as in Fig. 15.2. If we try to imagine how to determine $L_{12}$, it is not at all obvious because the surface of a toroidal coil is complicated. However, $L_{21} = \Phi_{21}/I_2$ is quite simple to find. The flux $d\Phi$ through the surface $dS = h\, dr$ in the figure is given by

$$d\Phi_{21}(r) = B(r)\, dS = \frac{\mu_0 N I_2}{2\pi r} h\, dr.$$

To obtain the total flux through $C_1$, we integrate over the cross section of the torus,

$$\Phi_{21} = \frac{\mu_0 N I_2 h}{2\pi} \int_a^b \frac{dr}{r} = \frac{\mu_0 N I_2 h}{2\pi} \ln\frac{b}{a},$$
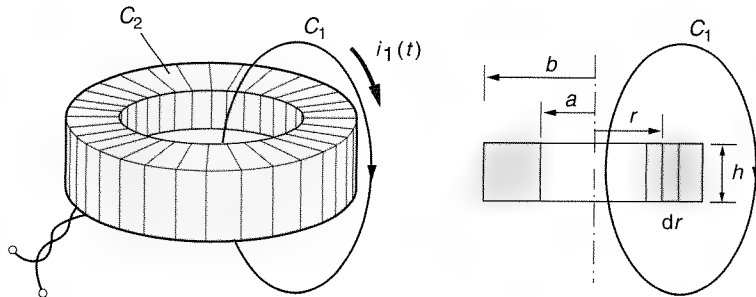


Figure 15.2 A toroidal coil and a single wire loop encircling the toroid

or

$$L_{12} = L_{21} = \frac{\mu_0 N h}{2\pi} \ln \frac{b}{a}.$$

Note that mutual inductance in this case does not depend at all on the shape of the wire loop. Also, if we need a larger mutual inductance (and thus larger induced emf), we can simply wind the loop two or more times around the toroid to obtain two or more times larger inductance. This is the principle of operation of transformers.

**Example 15.2—Mutual inductance of two coils wound on a toroidal core.** As another example, let us find the mutual inductance between two toroidal coils tightly wound one on top of the other on a core of the form shown in Fig. 15.2. Assume that one coil has $N_1$ turns and the other $N_2$ turns. If a current $I_2$ flows through coil 2, the flux through coil 1 is just $N_1$ times the flux $\Phi_{21}$ from the preceding example, where $N$ should be substituted by $N_2$. So

$$L_{12} = L_{21} = \frac{\mu_0 N_1 N_2 h}{2\pi} \ln \frac{b}{a}.$$

*Questions and problems:*   Q15.1 to Q15.7, P15.1 to P15.8

## 15.3   Self-Inductance

As mentioned in the introduction to this chapter, when a current in a contour varies in time, the induced electric field exists everywhere around it and therefore also along its entire length. Consequently there is an induced emf in the contour itself. This process is known as *self-induction*.

The simplest (but not physically the clearest) way of expressing this emf is to use Faraday's law:

$$e(t) = -\frac{d\Phi_{\text{self}}(t)}{dt}. \tag{15.6}$$

If the contour is in a linear medium (i.e., the flux through the contour is proportional to the current), we define the self-inductance of the contour as the ratio of the flux through the contour due to current $i(t)$ in it, and $i(t)$,

$$L = \frac{\Phi_{\text{self}}(t)}{i(t)} \quad (\text{H}). \tag{15.7}$$

Using this definition, the induced emf can be written as

$$e(t) = -L\frac{di}{dt}. \tag{15.8}$$

The constant $L$ depends only on the geometry of the system and the properties of the medium, and its unit is again a henry (H). In the case of a dc current, $L = \Phi/I$, which can be used for determining the self-inductance in some cases in a simple manner.

How do self-inductances of two contours compare with their mutual inductance? This is easy to answer for two simple loops. In that case, it is evident that the largest possible flux due to a current $i_1$ through a contour $C_1$ is the flux through the contour itself (the contour cannot be closer to any other contour than to itself). Therefore

$$\Phi_{11} \geq \Phi_{12} \quad \text{and} \quad \Phi_{22} \geq \Phi_{21}. \tag{15.9}$$

When we multiply these inequalities together and divide by $I_1 I_2$, we obtain

$$L_{11} L_{22} \geq L_{12}^2. \tag{15.10}$$

Therefore the largest possible value of mutual inductance is the geometric mean of the self-inductances. Although Eq. (15.10) is derived for a somewhat special case (two simple loops), it can be shown to be valid in general (see Example 16.2 in the following chapter).

Frequently, Eq. (15.10) is written as

$$L_{12} = k \sqrt{L_{11} L_{22}} \qquad -1 \leq k \leq 1. \tag{15.11}$$

The coefficient $k$ is called the *coupling coefficient*.

**Example 15.3—Self-inductance of a toroidal coil.** Consider again the toroidal coil in Fig. 15.2. If the coil has $N$ turns, what is its self-inductance?

In Example 15.1 we found the flux the coil produces through a cross section of the core. This flux exists through all the $N$ turns of the coil, so that the flux the coil produces through itself is simply $N$ times what we found in Example 15.1. The self-inductance of the coil in Fig. 15.2 is therefore

$$L = \frac{\mu_0 N^2 h}{2\pi} \ln\frac{b}{a}.$$

**Example 15.4—Self-inductance of a thin two-wire line.** Let us find the self-inductance per unit length of a thin two-wire line (Fig. 15.3). We can imagine that the line is actually a very long rectangular contour (closed with a load at one end and a generator at the other end), and that we are looking at only one part of it, hatched in the figure. At a distance $r$ from conductor 1, the current in it produces a magnetic flux density of intensity $B_1(r) = \mu_0 I/(2\pi r)$, and the current in conductor 2 a magnetic flux density $B_2(r) = \mu_0 I/[2\pi(d-r)]$. The total flux through a strip of width $dr$ and length $h$ shown in the figure is therefore

$$\Phi = \int_a^{d-a} [B_1(r) + B_2(r)] h \, dr = \frac{\mu_0 I h}{\pi} \ln\frac{d-a}{a} \simeq \frac{\mu_0 I h}{\pi} \ln\frac{d}{a},$$

since $d \gg a$. The inductance per unit length of the two-wire line is therefore

$$L' = \frac{\mu_0}{\pi} \ln\frac{d}{a}. \tag{15.12}$$
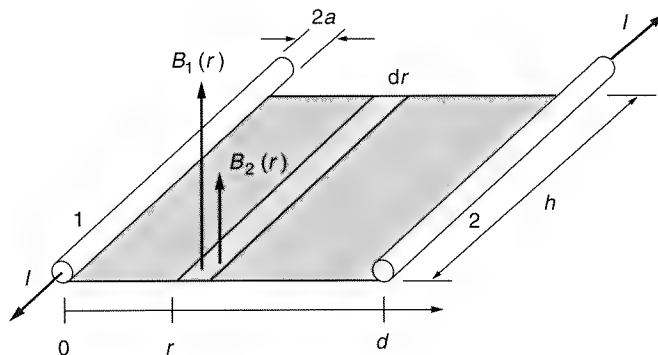
Figure 15.3 Calculating the self-inductance of a thin two-wire line

As a numerical example, for $d/a = 200$, $L' = 2.12\,\mu\text{H}/\text{m}$. We have only calculated the flux through the surface outside of the conductors. The expression for $L$ in Eq. (15.12) is therefore called the *external self-inductance* of the line. There is also an *internal self-inductance*, due to the flux through the wires themselves. We will introduce the concept of the internal inductance in terms of energy in the next chapter.

**Example 15.5—Self-inductance of a coaxial cable.** Let us find the external self-inductance per unit length of a coaxial cable. We first need to figure out through which surface to find the flux. If we imagine that the cable is connected to a generator at one end and to a load at the other, the current flows "in" through the inner conductor and flows back through the outer conductor. The flux through such a contour, for a cable of length $h$, is the flux through the rectangular surface in Fig. 15.4,
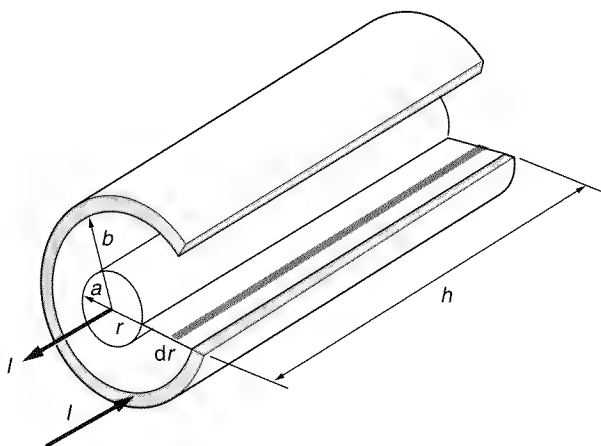


Figure 15.4 Calculating the self-inductance of a coaxial cable

$$\Phi = \int_a^b B(r)h\,dr = \frac{\mu_0 Ih}{2\pi}\ln\frac{b}{a}.$$

The external self-inductance per unit length of the cable is

$$L' = \frac{\mu_0}{2\pi}\ln\frac{b}{a}. \tag{15.13}$$

As a numerical example, for $b/a = e = 2.71828\ldots$, $L' = 0.2\,\mu\,H/m$. It is left as an exercise for the reader to calculate the inductance per unit length of the RG-55/U high-frequency coaxial cable from Example 8.7.

*Questions and problems:*   Q15.8 to Q15.20, P15.9 to P15.21

## 15.4   Chapter Summary

1. The coupling between two loops by means of the induced electric field is usually termed *magnetic coupling*.
2. The level of coupling between two loops is described by mutual inductance between the loops.
3. Mutual inductance can be calculated as $L_{12} = \Phi_{12}/I_1$, where $\Phi_{12}$ is the flux through contour 2 due to a current $I_1$ in contour 1.
4. A loop with a time-varying current produces an induced electric field also along the loop, which affects the current in the loop. This is known as self-induction, and the parameter describing it is self-inductance.
5. Self-inductance can be evaluated as the ratio of the flux through the contour due to a current in it, divided by that current ($L = \Phi/I$).

## QUESTIONS

**Q15.1.** What does the expression in Eq. (15.1) for the emf induced in a wire loop actually represent?

**Q15.2.** Why does mutual (and self) inductance have no practical meaning in the dc case?

**Q15.3.** Explain why mutual inductance for a toroidal coil and a wire loop encircling it (e.g., see Fig. P15.1) does not depend on the shape of the wire loop.

**Q15.4.** Explain in terms of the induced electric field why the emf induced in a coil encircling a toroidal coil and consisting of several loops (e.g., see Fig. P15.1) is proportional to the number of turns of the loop.

**Q15.5.** Can mutual inductance be negative as well as positive? Explain by considering reference directions of the loops.

**Q15.6.** Mutual inductance of two simple loops is $L_{12}$. We replace the two loops by two very thin coils of the same shapes, with $N_1$ and $N_2$ turns of very thin wire. What is the mutual inductance between the coils? Explain in terms of the induced electric field.

**Q15.7.** A two-wire line crosses another two-wire line at a distance $d$. The two lines are normal. Prove that the mutual inductance is zero, starting from the induced electric field.

**Q15.8.** In Example 15.3 we found that the self-inductance of a toroidal coil is proportional to the *square* of the number of turns of the coil. Explain this in terms of the induced electric field and induced voltage in the coil due to the current in the coil.

**Q15.9.** A thin coil is made of $N$ turns of very thin wire pressed tightly together. If the self-inductance of a single turn of wire is $L$, what do you expect is the self-inductance of the coil? Explain in terms of the induced electric field.

**Q15.10.** Explain in your own words what the meaning of self-inductance of a coaxial cable is.

**Q15.11.** Is it physically sound to speak about the mutual inductance between two wire segments belonging either to two loops or to a single loop? Explain.

**Q15.12.** Is it physically sound to speak about the self-inductance of a segment of a closed loop? Explain.

**Q15.13.** To obtain a resistive wire with the smallest self-inductance possible, the wire is sharply bent in the middle and the two mutually insulated halves are pressed tightly together, as shown in Fig. Q15.13. Explain why the self-inductance is minimal in terms of the induced electric field and in terms of the magnetic flux through the loop.
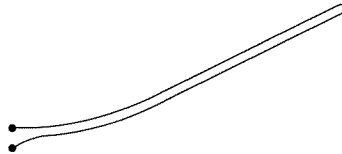
Figure Q15.13 A loop with small self-inductance

**Q15.14.** Can self-inductance be negative as well as positive? Explain in terms of the flux.

**Q15.15.** The self-inductance of two identical loops is $L$. What is approximately the mutual inductance between them if they are pressed together? Explain in terms of the induced electric field and in terms of the magnetic flux.

**Q15.16.** Two coils are connected in series. Does the total (equivalent) inductance of the connection depend on their mutual position? Explain.

**Q15.17.** Pressed onto a thin conducting loop is an identical thin *superconducting* loop. What is the self-inductance of the conducting loop? Explain.

**Q15.18.** A loop is connected to a source of voltage $v(t)$. As a consequence, a current $i(t)$ exists in the loop. Another conducting loop with no source is brought near the first loop. Will the current in the first loop be changed? Explain.

**Q15.19.** Answer question Q15.18 assuming that the source in the first loop is a dc source. Explain.

**Q15.20.** A thin, flat loop of self-inductance $L$ is placed over a flat surface of very high permeability. What is the new self-inductance of the loop?

## PROBLEMS

**P15.1.** Find the mutual inductance between an arbitrary loop and the toroidal coil in Fig. P15.1. There are $N$ turns around the torus, and the permeability of the core is $\mu$.
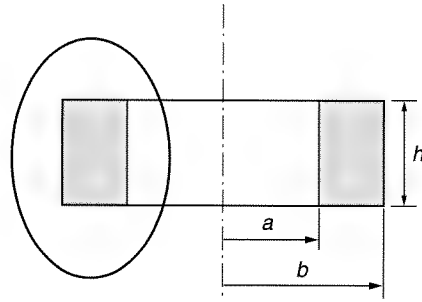
Figure P15.1 A toroidal coil and wire loop

**P15.2.** Find the mutual inductance of two two-wire lines running parallel to each other. The cross section of the lines is shown in Fig. P15.2.
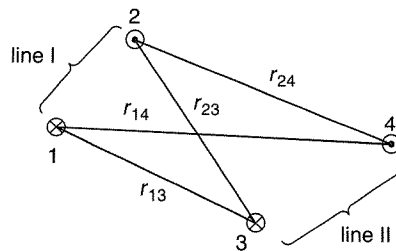


Figure P15.2 Two parallel two-wire lines

**P15.3.** A cable-car track runs parallel to a two-wire phone line, as in Fig. P15.3. The cable-car power line and track form a two-wire line. The amplitude of the sinusoidal current through the cable-car wire is $I_m$ and its angular frequency is $\omega$. All conductors are very
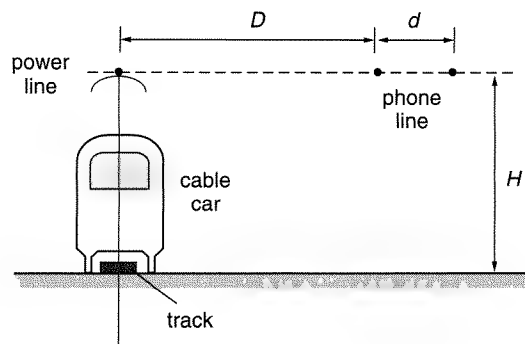


Figure P15.3 Cable-car track parallel to phone line

thin compared to the distances between them. Find the amplitude of the induced emf in a section of the phone line $b$ long.

**P15.4.** Parallel to a thin two-wire symmetrical power line along a distance $h$ is a thin two-wire telephone line, as shown in Fig. P15.4. (1) Find the mutual inductance between the two lines. (2) Find the amplitude of the emf induced in the telephone line when there is a sinusoidal current with amplitude $I_m$ and frequency $f$ in the power line. As a numerical example, assume the following: $f = 100\,\text{Hz}$, $I_m = 100\,\text{A}$, $h = 50\,\text{m}$, $d = 10\,\text{m}$, $a = 50\,\text{cm}$, and $b = 25\,\text{cm}$.
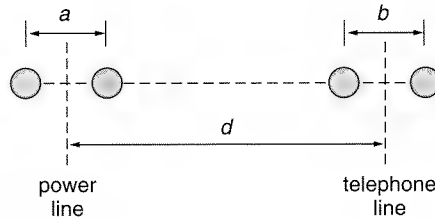


Figure P15.4 Parallel power and phone lines

**P15.5.** Two coaxial thin circular loops of radii $a$ and $b$ are in the same plane. Assuming that $a \gg b$ and that the medium is air, determine approximately the mutual inductance of the loops. As a numerical example, evaluate the mutual inductance if $a = 10\,\text{cm}$ and $b = 1\,\text{cm}$.

**P15.6.** Two coaxial thin circular loops of radii $a$ and $b$ are in air a distance $d$ ($d \gg a, b$) apart. Determine approximately the mutual inductance of the loops. As a numerical example, evaluate the mutual inductance if $a = b = 1\,\text{cm}$ and $d = 10\,\text{cm}$.

**P15.7.** Inside a very long solenoid wound with $N'$ turns per unit length is a small flat loop of surface area $S$. The plane of the loop makes an angle $\theta$ with the solenoid axis. Determine and plot the mutual inductance between the solenoid and the loop as a function of $\theta$. The medium is air.

**P15.8.** Assume that within a certain time interval the current in circuit 1 in Fig. P15.8 grows linearly, $i_1(t) = I_0 + It/t_1$. Will there be any current in circuit 2 during this time? If yes, what is the direction and magnitude of the current? The number of turns of the two coils is the same.
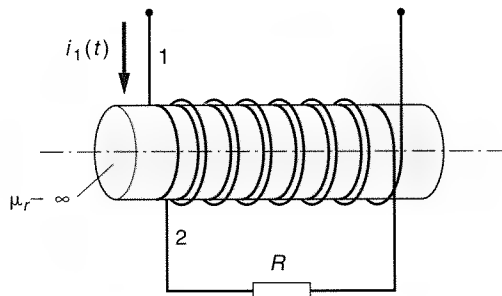


Figure P15.8 Two coupled circuits

**P15.9.** Three coupled closed circuits have self-inductances equal to $L_1$, $L_2$, and $L_3$, resistances $R_1$, $R_2$, and $R_3$, and mutual inductances $L_{12}$, $L_{13}$, and $L_{23}$. Write the equations for the currents in all three circuits if a voltage $v_1(t)$ is connected to circuit 1 only. Then write the equations for the case when three sources of voltages $v_1(t)$, $v_2(t)$, and $v_3(t)$ are connected to circuits 1, 2, and 3, respectively.

**P15.10.** A coaxial cable has conductors of radii $a$ and $b$. The inner conductor is coated with a layer of ferrite $d$ thick ($d < b - a$) and of permeability $\mu$. The rest of the cable is air-filled. Find the external self-inductance per unit length of the cable. What should your expression reduce to (1) when $d = 0$ and (2) when $d = b - a$?

**P15.11.** The conductor radii of a two-wire line are $a$ and the distance between them is $d$ ($d \gg a$). Both conductors are coated with a thin layer of ferrite $b$ thick ($b \ll d$) and of permeability $\mu$. The ferrite is an insulator. Calculate the external self-inductance per unit length of the line.

**P15.12.** The core of a toroidal coil of $N$ turns consists of two materials of respective permeabilities $\mu_1$ and $\mu_2$, as in each part of Fig. P15.12. Find the self-inductance of the toroidal coil and the mutual inductance between the coil and the loop positioned as in Fig. P15.1 if (1) the ferrite layers are of equal thicknesses, $h/2$, in Fig. P15.12a, and (2) the ferrite layers are of equal heights $h$ and the radius of the surface between them is $c$ ($a < c < b$), in Fig. P15.12b.
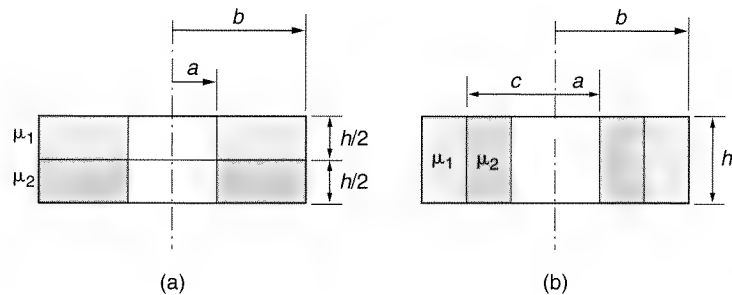


Figure P15.12 Two toroidal coils with inhomogeneous cores

**P15.13.** Three toroidal coils are wound in such a way that the coils 2 and 3 are inside coil 1, as in the cross section shown in Fig. P15.13. The medium is air. Find the self-inductances
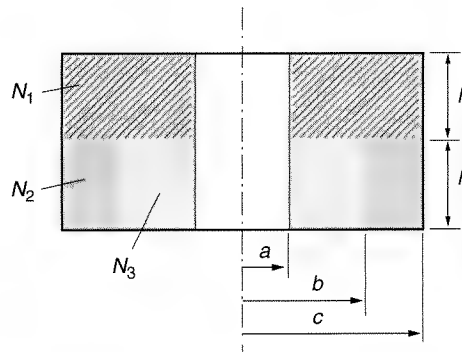


Figure P15.13 Three toroidal coils

$L_1$, $L_2$, and $L_3$ and mutual inductances $L_{12}$, $L_{13}$, and $L_{23}$. What are the different values of inductance that can be obtained by connecting the three windings in series in different ways?

**P15.14.** The width of the strips of a long, straight strip line is $a$ and their distance is $d$ (Fig. P15.14 for $d_2 = 0$). Between the strips is a ferrite of permeability $\mu$. Neglecting edge effects, find the inductance of the line per unit length.
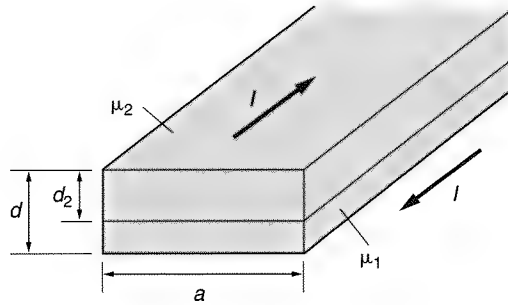


Figure P15.14 A strip line with a two-layer dielectric

**P15.15.** The width of the strips of a strip line is $a$ and their distance is $d$. Between the strips are two ferrite layers of permeabilities $\mu_1$ and $\mu_2$, and the latter is $d_2$ thick, as in Fig. P15.14. Neglecting edge effects, find the inductance of the line per unit length.

**P15.16.** A long thin solenoid of length $b$ and cross-sectional area $S$ is situated in air and has $N$ tightly wound turns of thin wire. Neglecting edge effects, determine the inductance of the solenoid.

**P15.17.** A thin toroidal core of permeability $\mu$, mean radius $R$, and cross-sectional area $S$ is densely wound with two coils of thin wire with $N_1$ and $N_2$ turns, respectively. The windings are wound one over the other. Determine the self- and mutual inductances of the coils and the coefficient of coupling between them.

**P15.18.** A thin toroidal ferromagnetic core of mean radius $R$ and cross-sectional area $S$ is densely wound with $N$ turns of thin wire. A current $i(t) = I_0 + I_m \cos \omega t$, where $I_0$ and $I_m$ are constants and $I_0 \gg I_m$, is flowing through the coil. Which permeability would you adopt in approximately determining the coil self-inductance? Assuming that this permeability is $\mu$, determine the self-inductance of the coil. Does it depend on $I_0$?

**P15.19.** A thin solenoid is made of a large number of turns of very thin wire tightly wound in several layers. The radius of the innermost layer is $a$, of the outermost layer $b$, and the solenoid length is $d$ ($d \gg a, b$). The total number of turns is $N$, and the solenoid core is made out of cardboard. Neglecting edge effects, determine approximately the solenoid self-inductance. Note that the magnetic flux through the turns differs from one layer to the next. Plot this flux as a function of radius, assuming the layers of wire are very thin.

**P15.20.** Repeat problem P15.19 for a thin toroidal core. Assume that the mean toroid radius is $R$, the total number of turns $N$, the radius of the innermost layer $a$, and that of the outermost layer $b$, with $R \gg a, b$.

**P15.21.** The current intensity in a circuit of self-inductance $L$ and negligible resistance was kept constant during a period of time at a level $I_0$. Then during a short time interval $\Delta t$, the current was linearly reduced to zero. Determine the emf induced in the circuit during this time interval. Does this have any connection with a spark you have probably seen inside a switch you turned off in the dark? Explain.

# 16

# Energy and Forces in the Magnetic Field

## 16.1  Introduction

Many devices make use of electric or magnetic forces. Most can be made in an electric version or a magnetic version. We shall see that magnetic forces are several orders of magnitude stronger than electric forces. Consequently, devices based on magnetic forces are much smaller and are used more often: for example, electric motors, large cranes for lifting ferromagnetic objects, doorbells, and electromagnetic relays.

This chapter derives the expressions for calculating magnetic energy, forces, and pressures. To a large extent, it parallels the chapter on electric energy, forces, and pressures, so the discussion is fairly brief.

## 16.2  Energy in the Magnetic Field

We did not mention energy when we discussed time-invariant magnetic fields because while establishing a dc current the current through a contour has to change from zero to its final dc value. During this process, there is a changing magnetic flux through the contour due to the changing current, and an emf is induced in the contour. This emf opposes the change of flux, according to Lentz's law. To establish the

278

final static magnetic field, the sources have to overcome this emf. Therefore, we could not talk about energy in the field without knowing about electromagnetic induction.

Let $n$ contours, with currents $i_1(t)$, $i_2(t)$, ..., $i_n(t)$, be the sources of a magnetic field. Assume that the contours have resistances $R_1, R_2, ..., R_n$ and are connected to generators of electromotive forces $e_1(t)$, $e_2(t)$, ..., $e_n(t)$. Finally, let the contours be stationary and rigid (i.e., they cannot be deformed), with *total* fluxes $\Phi_1(t)$, $\Phi_2(t)$, ..., $\Phi_n(t)$.

The work done by the voltage generators in *all* the contours during a short time interval $dt$ is partly converted into Joule's losses and partly used to change the magnetic field:

$$dA_g = dA_J + dA_m. \tag{16.1}$$

Generators in individual contours $k, k = 1, 2, ..., n$ do the following work:

$$(dA_g)_k = e_k(t)i_k(t)\,dt \qquad k = 1, 2, ..., n. \tag{16.2}$$

We also know that the total emf's in the loops are $e_k(t) - d\Phi_k(t)/dt$, so that $e_k(t) = R_k i_k(t) + d\Phi_k(t)/dt$. The last equation thus becomes

$$(dA_g)_k = R_k i_k^2(t)\,dt + i_k(t)\,d\Phi_k(t) \qquad k = 1, 2, ..., n. \tag{16.3}$$

The work of all the generators in the system is hence

$$dA_g = \sum_{k=1}^{n} R_k i_k^2(t)\,dt + \sum_{k=1}^{n} i_k(t)\,d\Phi_k(t). \tag{16.4}$$

The first term on the right-hand side is equal to the Joule's losses, $dA_J$, during time interval $dt$. Therefore, according to Eq. (16.1), the second term on the right-hand side is equal to $dA_m$ (the energy used to change the magnetic field):

$$dA_m = dA_g - dA_J = \sum_{k=1}^{n} i_k(t)\,d\Phi_k(t). \tag{16.5}$$

This equation expresses the law of conservation of energy for $n$ stationary current contours. $dA_m$ is the work necessary to change the fluxes through the $n$ contours by $d\Phi_1, d\Phi_2, ..., d\Phi_n$.

Let us now find the total work $A_m$ needed to establish dc currents $I_1, I_2, ..., I_n$ for which the fluxes through the contours are $\Phi_1, \Phi_2, ..., \Phi_n$. This is obtained by integrating the last equation from zero fluxes through the contours to fluxes $\Phi_1, \Phi_2, ..., \Phi_n$:

$$(A_m)_{\text{in establishing currents}} = \sum_{k=1}^{n} \int_0^{\Phi_k} i_k(t)\,d\Phi_k(t). \tag{16.6}$$

During the time needed to establish the fluxes $\Phi_k, k = 1, 2, ..., n$, the currents in the contours could have varied from zero to their final values in an infinite number of ways. From the law of conservation of energy, no matter how they have changed the final energy would have to be the same. If this is so, assume simply that all of the

currents changed linearly with time and that it took a time $T$ to establish the final dc currents. So we assume that $i_k(t) = I_k t/T$. Obviously the fluxes through the contours then also change linearly, $\Phi_k(t) = \Phi_k t/T$, so we have

$$(A_m)_{\text{in establishing currents}} = \sum_{k=1}^{n} \int_0^T I_k \frac{t}{T} \Phi_k \frac{dt}{T} = \sum_{k=1}^{n} \frac{1}{2} I_k \Phi_k. \tag{16.7}$$

This is valid only for linear media because we assumed that no work was spent on magnetizing any ferromagnetic body. The energy equal to this work is now stored in the magnetic field, and if the currents are reduced to zero this amount of energy is obtained from the system. Therefore we know that there is energy in a static magnetic field equal to

$$W_m = \frac{1}{2} \sum_{k=1}^{n} I_k \Phi_k. \tag{16.8}$$

*(Magnetic energy of n current contours)*

This can also be expressed in terms of self- and mutual inductances of the contours and currents in them. First, the *total* flux through the $k$-th contour (due to the current in itself and in all the other contours) can be expressed as

$$\Phi_k = \Phi_{1k} + \Phi_{2k} + \cdots + \Phi_{kk} + \cdots + \Phi_{nk}$$
$$= L_{1k}I_1 + L_{2k}I_2 + \cdots + L_{kk}I_k + \cdots + L_{nk}I_n. \tag{16.9}$$

This can be written in the form

$$\Phi_k = \sum_{j=1}^{n} L_{jk}I_j. \tag{16.10}$$

Thus the magnetic energy in Eq. (16.8) of $n$ contours with currents $I_1, I_2, \ldots, I_n$ can also be written as

$$W_m = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} L_{jk}I_jI_k. \tag{16.11}$$

*(Magnetic energy of n current contours)*

The important case is that of a single contour:

$$W_m = \frac{1}{2} I\Phi = \frac{1}{2} LI^2. \tag{16.12}$$

*(Magnetic energy of a single current contour)*

**Example 16.1—Magnetic energy of two magnetically coupled contours.** In the case of two contours ($n = 2$), Eqs. (16.8) and (16.11) for the magnetic energy of $n$ contours become

$$W_m = \frac{1}{2}(I_1\Phi_1 + I_2\Phi_2)$$

and

$$W_m = \frac{1}{2}L_{11}I_1^2 + \frac{1}{2}L_{22}I_2^2 + L_{12}I_1I_2.$$

This energy can be smaller or larger than the sum of energies of the two contours when isolated because $L_{12}$ can be positive or negative.

**Example 16.2—General proof that $|L_{12}| \leq \sqrt{L_{11}L_{22}}$.** We have proved the inequality $|L_{12}| \leq \sqrt{L_{11}L_{22}}$ considering two simple loops only. We now prove that this is true for any two contours.

If $I_1$ is kept constant, the preceding equation can be rewritten as

$$W_m = I_1^2 \left( \frac{1}{2}L_{11} + \frac{1}{2}L_{22}x^2 + L_{12}x \right), \qquad \text{where} \qquad x = I_2/I_1.$$

The magnetic energy $W_m$ is always larger than zero. Its minimum is found from $dW_m/dx = 0$:

$$\frac{dW_m}{dx} = I_1^2(L_{22}x + L_{12}) = 0.$$

The latter is true for $x = -L_{12}/L_{22}$. For this value of $x$, the expression for the magnetic energy becomes

$$(W_m)_{min} = \frac{1}{2}I_1^2(L_{11} - L_{12}^2/L_{22}).$$

Because $W_m \geq 0$, we see that for any two coupled contours, $L_{12}^2 \leq L_{11}L_{22}$.

*Questions and problems:*   Q16.1 to Q16.15, P16.1

## 16.3   Distribution of Energy in the Magnetic Field

We saw earlier that in the electrostatic field we could find the energy in two ways: as a potential energy of a system of charges or as energy distributed in the entire field with a certain density. We shall now show that an expression of the form in Eq. (9.7) can also be derived for the energy of a magnetic field.

Consider first a simple example, a thin torus of radius $R$ and core cross-sectional area $S$ with $N$ turns carrying a current $i(t)$. The core can be of any homogeneous magnetic material. The magnetic field in the torus is

$$H(t) = \frac{Ni(t)}{2\pi R}, \tag{16.13}$$

from which $i(t) = 2\pi RH(t)/N$. Let $d\Phi(t) = S\,dB(t)$ be the increase in the flux in the core of the torus during a short time interval $dt$. Then the increase in the flux in all the $N$ turns is $NS\,dB(t)$. According to the formula in Eq. (16.6), the work done by the sources to change the flux through the torus from $\Phi_1$ to $\Phi_2$ is

$$(A_m)_{\text{from } \Phi_1 \text{ to } \Phi_2} = \int_{\Phi_1}^{\Phi_2} i(t)\,d\Phi(t) = 2\pi RS \int_{B_1}^{B_2} H(t)\,dB(t), \tag{16.14}$$

where $B_1$ is the initial magnetic flux density and $B_2$ the final flux density. Because $2\pi RS$ is the volume of the torus, we see that the volume energy density spent in order to change the magnetic flux density vector from $B_1$ to $B_2$ is equal to

$$\frac{dA_m}{dv} = \int_{B_1}^{B_2} H(t)\,dB(t). \tag{16.15}$$

*(Density of work that needs to be done to change B from $B_1$ to $B_2$ at a point)*

This formula was derived for a special case of a toroidal coil. It can be shown that it remains valid for an arbitrary magnetic field (in a manner analogous to that which we used for the electric energy, when we generalized the proof obtained for a parallel-plate capacitor to the general case). In the present case, we imagine the entire field divided into elemental tubes of flux of vector **B**. The proof, which is somewhat more complicated than in the electrostatic case but quite analogous, is left to the interested reader as an exercise.

In the case of linear media, energy used for changing the magnetic field is stored in the field, that is, $dA_m = dW_m$. Assuming that the $B$ field changed from zero to some value $B$, we have

$$\frac{dW_m}{dv} = \int_0^B \frac{B}{\mu}\,dB = \frac{1}{2}\frac{B^2}{\mu} = \frac{1}{2}\mu H^2 = \frac{1}{2}BH. \tag{16.16}$$

*(Density of energy in magnetic field—linear media only)*

The energy *in a linear medium* can now be found by integrating over the entire volume of the field:

$$W_m = \int_v \frac{1}{2}\mu H^2 dv. \tag{16.17}$$

*(Magnetic energy distributed over the entire field—linear media only)*

**Example 16.3—Losses in ferromagnetics due to hysteresis.** Let us observe what happens to energy spent in maintaining a sinusoidal magnetic field in a piece of ferromagnetic material. The hysteresis curve of the material is shown in Fig. 16.1, and the arrows show the direction in which the point describing the curve is moving in the course of time. According
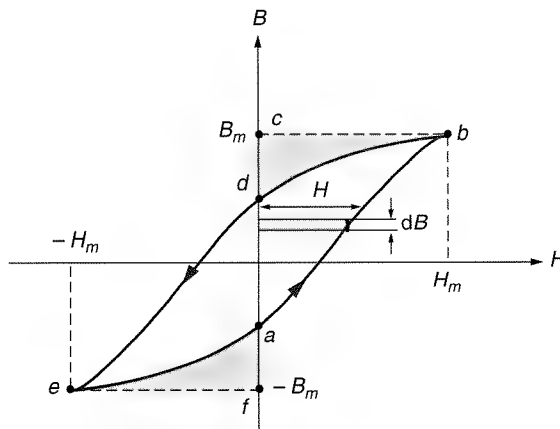
Figure 16.1 Hysteresis curve of a ferromagnetic material

to Eq. (16.15), the energy density that needs to be spent at a point where the magnetic field is $H$, in order to change the magnetic flux density by $dB$, is equal to $H\,dB$. In the diagram in Fig. 16.1, this is proportional to the area of the little shaded rectangle. So the integral of $H\,dB$ is proportional to the sum of all such rectangles as the point moves around the hysteresis curve.

Let us start from point $a$ in the figure. From $a$ to $b$, the magnetic field $H$ is positive. The increase $dB$ is also positive, so $H\,dB$ is positive and the energy density spent moving from point $a$ to $b$ is proportional to the area of the curved triangle $abc$ in the figure.

From $b$ to $d$, $H$ is positive but $B$ is decreasing, so that $dB$ is negative. Therefore the product $H\,dB$ is negative, which means that in this region the energy spent on maintaining the field is negative. This in turn means that this portion of the energy is returned from the field to the sources. The density of this returned energy is proportional to the area of the curved triangle $bdc$.

From $d$ to $e$ the product $HdB$ is positive, so this energy is spent on maintaining the field, and from $e$ to $a$ the product is negative, so this energy is returned to the sources. So only the energy density proportional to the area of the curved triangles $bcd$ and $efa$ is returned to the sources. All the rest, which is proportional to the area formed by the hysteresis loop, is lost to heat in the ferromagnetic material. These losses are known as *hysteresis losses*. If the frequency of the field is $f$, the loop is circumscribed $f$ times per second. Consequently, *the power of hysteresis losses is proportional to frequency* (and to the volume of the ferromagnetic material if the field is uniform).

**Example 16.4—Internal inductance of a straight wire.** The energy of a wire with a current $i$ is distributed around the wire as well as inside the wire because there is a magnetic field both outside and inside the wire. From the energy expression $W_m = \frac{1}{2}Li^2$ for a single current contour, we can write

$$L_{\text{internal}} = \frac{2(W_m)_{\text{inside conductor}}}{i^2}$$

and

$$L_{\text{external}} = \frac{2(W_m)_{\text{outside conductor}}}{i^2}.$$
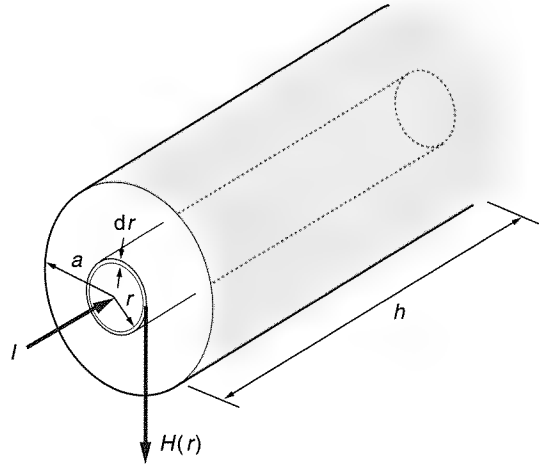
Figure 16.2 A long straight wire of circular cross section

A long straight wire of circular cross section is shown in Fig. 16.2. According to Ampère's law, the magnetic field inside the wire is equal to $H(r) = (Ir)/(2\pi a^2)$, so the energy density in the wire is

$$\frac{\mathrm{d}W_m}{\mathrm{d}v} = \frac{1}{2}\mu H^2 = \frac{1}{2}\frac{\mu I^2 r^2}{(2\pi a^2)^2}.$$

The magnetic energy stored in a length $h$ of the wire is obtained by integrating the last expression over the volume of the wire segment. The integration is easily done if the volume element is chosen to be a thin tube shown in the figure. The volume of the tube is $\mathrm{d}v = 2\pi rh\,\mathrm{d}r$. We thus find

$$(W_m)_{\text{inside wire}} = \int_0^a \frac{1}{2}\mu H^2(r)2\pi rh\,\mathrm{d}r = \frac{\mu I^2 h}{4\pi a^4}\int_0^a r^3\mathrm{d}r = \frac{\mu I^2 h}{16\pi}.$$

The internal self-inductance per unit length of a wire is hence

$$L'_{\text{internal}} = \frac{\mu}{8\pi}. \tag{16.18}$$

Note that the internal inductance does not depend on the radius of the wire. (Also note that it is very difficult to find internal inductance using the methods from Chapter 15 for external self- and mutual inductance. Why?)

**Example 16.5—Total inductance of a thin two-wire line.** Let us find the total self-inductance per unit length of a thin two-wire line with wires made of a material with permeability $\mu$. Assume that the wire radius is $a$ and the distance between the wire axes $d$. We found the external inductance in Example 15.4, so

$$L' = L'_{\text{external}} + L'_{\text{internal}} = \frac{\mu_0}{\pi}\ln\frac{d}{a} + 2\frac{\mu}{8\pi}. \tag{16.19}$$

We multiplied the expression for the internal inductance of a wire by 2 because there are two wires in the line. As a numerical example, if $\mu = \mu_0$ and $d/a = 100$, we get $L'_{\text{external}} = 1.84\,\mu\text{H/m}$ and $L'_{\text{internal}} = 0.1\,\mu\text{H/m}$. In this example, the external inductance is much larger than the internal inductance. This is usually the case.

*Questions and problems:*   Q16.16 to Q16.29, P16.2 to P16.15

## 16.4   Magnetic Forces

Suppose we know the distribution of currents, and that the currents exist in a magnetically homogeneous medium. In this case, the Biot-Savart law can be used for determining the magnetic flux density. Combined with the relation $d\mathbf{F}_m = I\,d\mathbf{l} \times \mathbf{B}$, we can find the magnetic force on any part of the current distribution. In many cases, however, this is quite complicated.

Similarly to finding the electric force from a change in energy, we can find the magnetic force as a derivative of the magnetic energy. This can be done assuming either (1) the fluxes through all the contours are kept constant or (2) the currents in all the contours are kept constant.

Assume first that during a displacement $dx$ of a body in the magnetic field along the $x$ axis we keep the fluxes through all the contours constant. This, of course, can be done by varying the currents in the contours appropriately. According to Eq. (16.5), during such a displacement the sources do not perform any work. (In fact, this situation corresponds to all the loops being superconducting, when no change of flux is possible—see Example 14.7.) Therefore, the work by the magnetic force in moving the body was done at the expense of the magnetic energy of the system:

$$F_x = -\left(\frac{dW_m}{dx}\right)_{\Phi=\text{constant}}. \tag{16.20}$$

In the second case, when the currents are kept constant, the fluxes can change. Therefore the sources have to do some work during the displacement, and it can be shown that

$$F_x = +\left(\frac{dW_m}{dx}\right)_{I=\text{constant}}. \tag{16.21}$$

**Example 16.6—Lifting force of an electromagnet.** As an example of the first formula let us find the attractive force of an electromagnet, sketched in Fig. 16.3. The electromagnet is in the shape of a horseshoe and its magnetic force is lifting a weight $W$, shown in the figure. This is a magnetic circuit. Let us assume that when the weight $W$ moves by a small amount $dx$ upward, the flux in the magnetic circuit does not change. That means that when the weight is moved upward the only change in magnetic energy is the reduction in energy contained in the two air gaps due to their decreased length. This energy reduction is

$$-dW_m = \frac{1}{2}\frac{B^2}{\mu_0}2S\,dx,$$
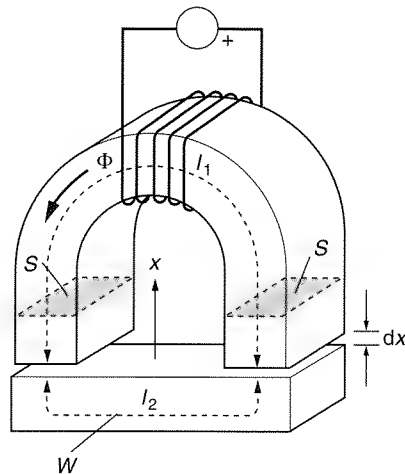
Figure 16.3 An electromagnet

so that

$$F_x = \frac{1}{2}\frac{B^2}{\mu_0}2S = \frac{\Phi^2}{\mu_0 S}.$$

As a numerical example, let $B = 1\,\mathrm{T}$ and $S = 1000\,\mathrm{cm}^2$. For this case, $F_x = 7.96 \cdot 10^4\,\mathrm{N}$, which means that this electromagnet can lift a weight of about 8 tons! Such electromagnets are used in cranes for lifting large pieces of iron, for example.

**Example 16.7—Magnetic force acting on a rectangular loop in the field of a straight wire with current.** As an example of the second formula for calculating the magnetic forces, consider a rectangular contour with a current $I_2$ that is $y$ away from a long straight wire with current $I_1$, as shown in Fig. 16.4. Let us find all three components of the force, $F_x$, $F_y$, and $F_z$.

There is obviously no flux change through the contour if it is moved in the $x$ or $z$ directions. So $F_x = F_z = 0$, and we need to find only $F_y$. When the currents through the wires are kept constant, according to the last equation in Example 16.1 we can write

$$F_y = \frac{\mathrm{d}W_\mathrm{m}}{\mathrm{d}y} = \frac{\mathrm{d}(L_{12}I_1I_2)}{\mathrm{d}y} = I_2\frac{\mathrm{d}\Phi_{12}}{\mathrm{d}y}, \qquad I_2 \text{ is constant.}$$

The change of flux through the rectangular contour is only due to the current in the straight wire,

$$\Phi_{12}(y) = \frac{\mu_0 I_1 b}{2\pi}\ln\frac{y+a}{y}.$$
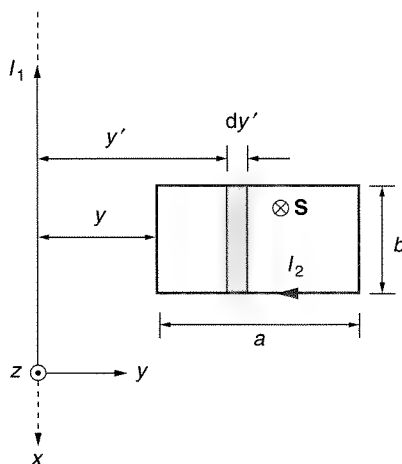
Figure 16.4 An example of the magnetic force calculation

so that

$$F_y = I_2 \frac{\mathrm{d}\Phi_{12}}{\mathrm{d}y} = -\mu_0 \frac{I_1 I_2 b}{2\pi} \frac{a}{y(y+a)}.$$

The negative sign means that the force is in the $-y$ direction, i.e., it is attractive.

**Example 16.8—An ammeter.** A possible way to build a simple ammeter using magnetic forces is shown in Fig. 16.5. A piece of ferromagnetic material, for example an iron nail (of cross-sectional radius $a$), is inserted partway into a solenoid and hangs off a spring. The relative
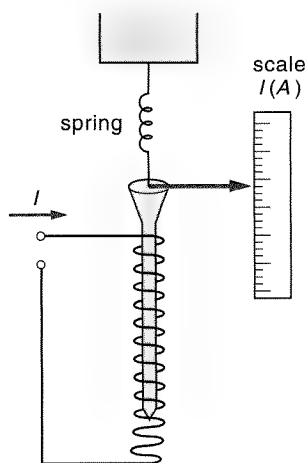


Figure 16.5 A simple ammeter

position of the nail in the vertical direction can be measured against a scale. When no current is flowing through the solenoid, the nail position is at zero. When a current flows through the coil, there is a force acting on the nail in addition to the gravitational force, the nail moves downward, and the new position of the nail is a direct (but not linear) measure of the current intensity in the coil. As an exercise (see problem P16.29), plot the "scale calibration curve" of such an ammeter. For what current levels is it useful given the dimensions in P16.29?

**Example 16.9—Comparison of electric and magnetic pressure.** We derived the expressions for the pressure of electric forces starting from the formulas analogous to Eqs. (16.20) and (16.21). The derivation of pressure in this case is therefore completely the same and will not be repeated here (although it is suggested to the reader to repeat the derivation as an exercise). For two magnetic media of permeabilities $\mu_1$ and $\mu_2$, the pressure on the interface, assumed to be directed into medium 1, is given by

$$p = \frac{1}{2}(\mu_2 - \mu_1)\left(H_{\text{tang}}^2 + \frac{B_{\text{norm}}^2}{\mu_1\mu_2}\right) \qquad \text{(reference direction of pressure into medium 1).}$$

$$(16.22)$$

We know that magnetic flux density of about $1\,\text{T}$ is quite large and not easily attainable. Therefore, according to the expression derived in Example 16.6, the maximal magnetic pressure that can be obtained is on the order of

$$(p_m)_{\text{max}} = \frac{1^2}{2 \cdot 4\pi \cdot 10^{-7}} \simeq 400,000\,\frac{\text{N}}{\text{m}^2}.$$

The electric pressure on a metallic conductor in a vacuum is given in Eq. (9.18), which can be rewritten as $p_e = \frac{1}{2}\epsilon_0 E^2$. We know that the electric strength of air is about $3 \cdot 10^6\,\text{V/m}$. This means that the largest electric pressure in air is approximately

$$(p_e)_{\text{max}} = 0.5 \cdot 8.86 \cdot 10^{-12} \cdot (3 \cdot 10^6)^2 \simeq 40\,\frac{\text{N}}{\text{m}^2}.$$

Consequently, the ratio of the maximal magnetic and maximal electric pressure is approximately

$$\frac{(p_m)_{\text{max}}}{(p_e)_{\text{max}}} = 10,000.$$

This is an extremely important conclusion. Although we can have electric and magnetic versions of almost any device using electric and magnetic forces, the magnetic version will require much less space for the same amount of power.

*Questions and problems:*    Q16.30 and Q16.31, P16.16 to P16.30

# 16.5   Chapter Summary

1. The energy necessary for establishing a magnetic field can be calculated in two ways: in terms of currents in wire loops or as an integral of energy density over the entire field.

2. If there are no losses (such as hysteresis losses), the energy used for creating the magnetic field can be retrieved when the field is switched off, so it represents the energy of the magnetic field.

3. The power of hysteresis losses is proportional to the area of the hysteresis loop and to frequency.

4. The energy concept of self-inductance indicates that it can be represented as a sum of the energy associated with the field external to the region with the current (the external inductance) and that associated with the field inside the current region (the internal inductance).

5. For current loops in a vacuum, the magnetic force on any loop can always be calculated, but this may be difficult. The energy-based approach to calculating magnetic forces in such cases might be simpler. In particular, if magnetic materials are present, magnetic forces can be evaluated by formulas based on the law of conservation of energy in the magnetic field.

# QUESTIONS

**Q16.1.** What does Eq. (16.1) actually represent?

**Q16.2.** Explain why the expression $dA_g = e(t)\, i(t)\, dt$ is the work done by a generator.

**Q16.3.** For a simple circuit of resistance $R$, with an emf $e(t)$, $e(t) = Ri(t) + d\Phi(t)/dt$. Explain the physical meaning of the last term.

**Q16.4.** Why does the energy of a system of current loops not depend on how the currents in the loops attained their final values?

**Q16.5.** Is Eq. (16.11) valid for nonlinear magnetic media? Explain.

**Q16.6.** The current in a thin loop 1 is increased from zero to a constant value $I$. A thin resistive loop 2 has no generators in it, but is in the magnetic field of the current in loop 1. Both loops are made of a linear magnetic material. Are the power $p_g(t)$ of the generator in loop 1 and the final value $W_m$ of the energy *stored* in the system affected by the presence of loop 2?

**Q16.7.** Repeat question Q16.6 with loop 2 open-circuited.

**Q16.8.** A body of a linear magnetic material is placed in the vicinity of loop 1 of question Q16.6. Is some energy associated with the magnetization of the body?

**Q16.9.** Equation (16.8) was derived by assuming that the currents were increased inside *stationary* conductors. Using the law of conservation of energy as an argument, prove that this expression must be valid for the magnetic energy of the system considered, irrespective of the process by which the current system is obtained.

**Q16.10.** Using a sound physical argument, explain why the work in Eq. (16.6) done by the generators in establishing a given time-constant magnetic field is a function of the process by which the system of currents is established when ferromagnetic materials are present in the field.

**Q16.11.** Will the magnetic energy of a system of fixed quasi-filamentary dc current loops be changed if a closed conducting loop with *no* current is introduced into the system? Explain.

**Q16.12.** Is it possible to determine theoretically the self-inductances and mutual inductances in a system of current loops by starting from Eq. (16.11) if $W_m$ is known? Explain.

**Q16.13.** Two equal thin loops of self-inductance $L$ are pressed onto each other so that $|L_{12}| \simeq L$. If the currents in the loops are $I_1$ and $I_2$, what is the magnetic energy of the system? Answer the question if the two currents are (1) in the same direction and (2) in opposite directions.

**Q16.14.** How would you make an electric version of a generator of sinusoidal emf?

**Q16.15.** How would you make an electric version of a generator of "rectified" sinusoidal emf?

**Q16.16.** Imagine somebody came to you with a piece of a ferromagnetic material he developed, and stated that the working point moves along the hysteresis loop in the clockwise direction. Would you believe him? Explain.

**Q16.17.** Is it possible to derive Eq. (16.15) from Eq. (16.16)? Explain.

**Q16.18.** If a hysteresis loop was obtained by a sinusoidally varying $H(t)$, will the hysteresis losses be exactly equal to the area of this loop if $H(t)$ varies as a triangular function of time (i.e., varies linearly from $-H_m$ to $H_m$, then back to $-H_m$, and so on)? Explain.

**Q16.19.** If the frequency of the alternating current producing a magnetic field is $f$ (cycles per second), what is the power per unit volume necessary to maintain the field in a piece of ferromagnetic substance?

**Q16.20.** According to the expression in Eq. (16.16), the volume density of magnetic energy is always greater in a vacuum than in a paramagnetic or idealized linear ferromagnetic material for the same $B$. Using a sound physical argument, explain this result.

**Q16.21.** The magnetization curve of a real ferromagnetic material is approximated by a nonlinear, but single-valued, function $B(H)$ (not by a hysteresis loop). Is it possible to speak about the energy density of the magnetic field inside the material? If you think it is, what is the energy density equal to?

**Q16.22.** A thin toroidal ferromagnetic core is magnetized to saturation and the current in the excitation coil is reduced to zero, so that the operating point in the $H$-$B$ plane is $H = 0$, $B = B_r$. Is it possible to speak in that case about the energy of the magnetic field stored in the core? Is it possible to speak about the energy used to create the field? Explain.

**Q16.23.** The first part of the magnetization curve can be approximated as $B(H) = CH^2$, where $C$ is a constant. How can you evaluate in that case the energy density necessary for the magnetization of the material? Is that also the energy density of the magnetic field?

**Q16.24.** Is it possible for the initial magnetization curve to be partly decreasing in $B$ as $H$ increases? What would that mean?

**Q16.25.** Evaluate approximately the density of hysteresis losses per cycle for the hysteresis loop in Fig. Q16.25 if $H_m = 200\,\text{A}/\text{m}$ and $B_m = 0.5\,\text{T}$.

**Q16.26.** Why are hysteresis losses linearly proportional to frequency?

**Q16.27.** Is the volume density of hysteresis losses in a thick toroidal ferromagnetic core with a coil carrying a sinusoidal current the same at all points of the core? Is the answer the same if the current intensity is such that at all points of the core saturation is attained, and if it is not?

**Q16.28.** Why can the self-inductance of a thick conductor not be defined naturally in terms of the induced emf or flux through the conductor?

**Q16.29.** Why is it very difficult to obtain the internal inductance using the methods from Chapter 15 for mutual inductance and self-inductance? To answer the question, con-
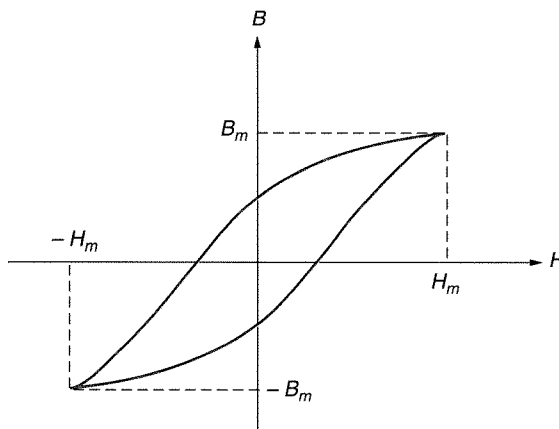
Figure Q16.25 A hysteresis loop

sider two wires, one thin and the other thick, with the same current $I$ flowing through them.

**Q16.30.** Direct current due to a lightning stroke on a three-phase line propagates along the three conductors. Will the force repel or attract the conductors?

**Q16.31.** If a lightning stroke hits a transformer, in some cases it may be noticed that the transformer "swells" (increases in volume). Explain why.

# PROBLEMS

**P16.1.** Write the explicit expression for the magnetic energy of three current loops with currents $I_1$, $I_2$, and $I_3$. Assume that the self-inductances and mutual inductances of the loops are known.

**P16.2.** Find the magnetic energy per unit length in the dielectric of a coaxial cable of inner conductor radius $a$ and outer conductor radius $b$, carrying a current $I$. The permeability of the dielectric is $\mu_0$. Show that $W'_m = L'_{\text{external}} I^2 / 2$.

**P16.3.** Find the total inductance per unit length of a coaxial cable that has an inner conductor of radius $a$ and an outer conductor with inner radius $b$ and outer radius $c$. The permeability of the conductors and the dielectric is $\mu_0$, and current is distributed uniformly over the cross sections of the two conductors.

**P16.4.** A thin ferromagnetic toroidal core is made of a material that can be characterized approximately by a constant permeability $\mu = 4000\mu_0$. The mean radius of the core is $R = 10$ cm and the core cross-sectional area is $S = 1$ cm$^2$. A current of $I = 0.1$ A is flowing through $N = 500$ turns wound around the core. Find the energy spent on magnetizing the core. Is this equal to the energy contained in the magnetic field in the core?

**P16.5.** In the toroidal core of the preceding problem, a small part of length $l_0 = 2$ mm is cut out so that now there is a small air gap in the core. The current in the coil is kept

constant while the piece is being cut out. Find the energy contained in the magnetic field in this case.

**P16.6.** Show that the same expression for the self-inductance of the toroidal coil in Fig. P16.6, as calculated in Example 15.3, is obtained from the expression $2W_m = LI^2$.
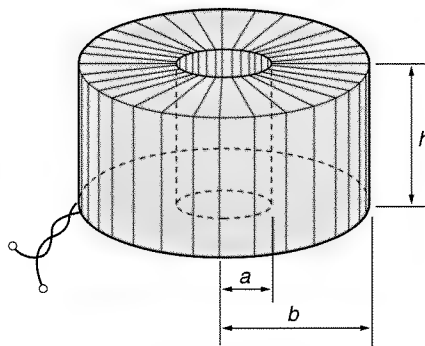


Figure P16.6 A thick toroidal coil

**P16.7.** On a thin ferromagnetic toroidal core of cross-sectional area $S = 1\,\text{cm}^2$, $N = 1000$ turns of thin wire are tightly wound. The mean radius of the core is $R = 16\,\text{cm}$. It may be assumed that the magnetic field is uniform over the cross section of the toroid. The idealized initial magnetization curve is shown in Fig. P16.7. Determine the work $A_m$ done in establishing the magnetic field inside the toroid if the intensity of the current through the coil is $I = 2\,\text{A}$.
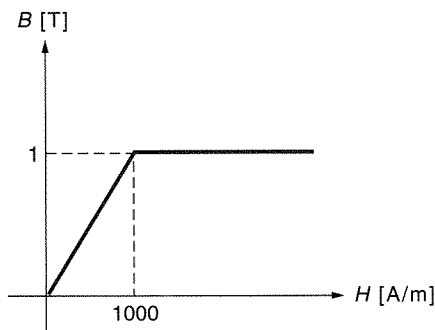


Figure P16.7 An idealized magnetization curve
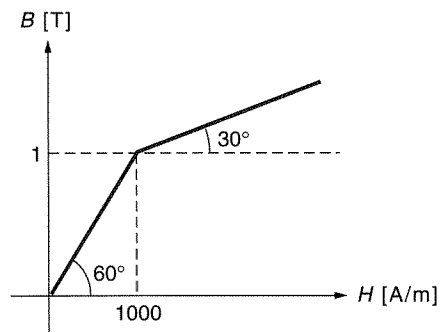


Figure P16.8 An idealized magnetization curve

**P16.8.** Repeat problem P16.7 if the idealized initial magnetization curve is as shown in Fig. P16.8.

**P16.9.** On the toroidal core shown in Fig. P16.6, $N = 650$ turns of thin wire are tightly wound. The intensity of the time-constant current in the coil is $I = 2\,\text{A}$ and the idealized initial magnetization curve of the core is as shown in Fig. P16.7. Determine the work done in establishing the magnetic field in the core if $a = 5\,\text{cm}$, $b = 15\,\text{cm}$, and $h = 10\,\text{cm}$.

**P16.10.** Repeat the preceding problem for intensities of current through the coil of (1) 0.5 A and (2) 1 A.

**P16.11.** The initial magnetization curve of a ferromagnetic material can be approximated by $B(H) = B_0H/(H_0 + H)$, where $B_0$ and $H_0$ are constants. Determine the work done per unit volume in magnetizing this material from zero to a magnetic field intensity $H$.

**P16.12.** The idealized hysteresis loops of the ferromagnetic core in Fig. P16.6 are as in Fig. P16.12. Determine the power of hysteresis losses in the core if it is wound with $N$ turns of wire with sinusoidal current of amplitude $I_m$ and frequency $f$. Assume that saturation is not reached at any point, and neglect the field of eddy currents.
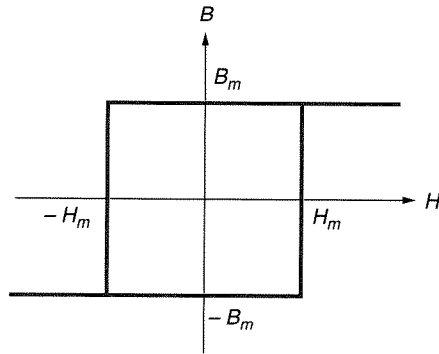


Figure P16.12 An idealized hysteresis loop    Figure P16.13 An idealized hysteresis loop
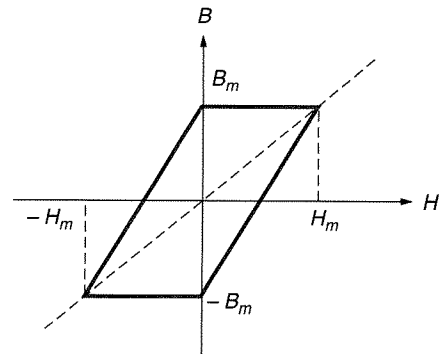
**P16.13.** Repeat problem P16.12 for idealized hysteresis loops shown in Fig. P16.13, assuming $B_m/H_m$ for all the loops is the same and that saturation is not reached at any point. Ignore the field of eddy currents.

**P16.14.** A ferromagnetic core of a solenoid is made of thin, mutually insulated sheets. To estimate the eddy current and hysteresis losses, the total power losses were measured at two frequencies, $f_1$ and $f_2$, for the same amplitude of the magnetic flux density. The total power losses were found to be $P_1$ and $P_2$, respectively. Determine the power of hysteresis losses and of eddy current losses at both frequencies.

***P16.15.** Prove that Eq. (16.15) is valid for any magnetic field, not necessarily uniform.

**P16.16.** Two coaxial solenoids of radii $a$ and $b$, lengths $l_1$ and $l_2$, and number of turns $N_1$ and $N_2$ have the same current $I$ flowing through them. Find the axial force that the solenoids exert on each other if the thinner solenoid is pulled by a length $x$ ($x < l_1, l_2$) into the other solenoid. Neglect edge effects and assume that the medium is air.

**P16.17.** An electromagnet and the weight it is supposed to lift are shown in Fig. P16.17. The dimensions are $S = 100 \, \text{cm}^2$, $l_1 = 50 \, \text{cm}$, $l_2 = 20 \, \text{cm}$. Find the current through the winding of the electromagnet and the number of turns in the winding so that it can lift a load that is $W = 300 \, \text{kiloponds}$ (a kp is 9.81 N) heavy. The electromagnet is made of a material whose magnetization curve can be approximated by $B(H) = 2H/(400+H)$, where $B$ is in T and $H$ is in A/m.
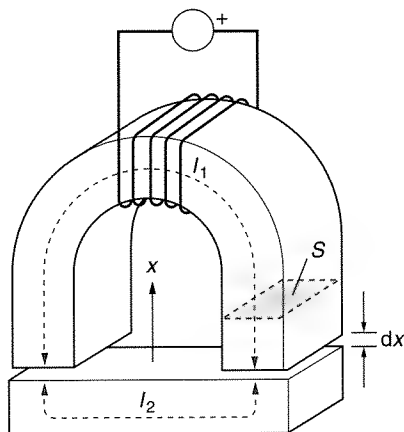
Figure P16.17 An electromagnet
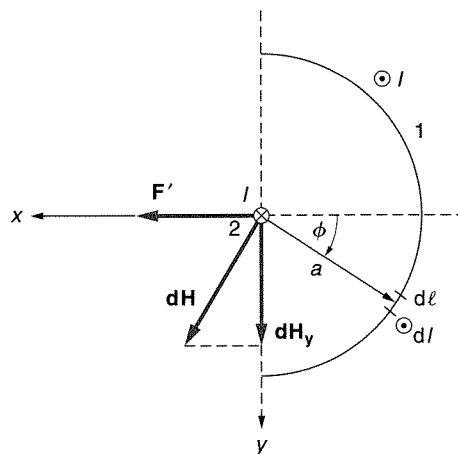
Figure P16.18 Cross section of a
two-conductor line

**P16.18.** One of the conductors of a two-conductor line is in the form of one half of a thin circular cylinder. The other conductor is a thin wire running along the axis of the first (Fig. P16.18). If a current $I$ flows through the two conductors in opposite directions, determine the force per unit length on the conductors.

**P16.19.** A thin conductor 2 runs parallel to a thin metal strip 1 (Fig. P16.19). Both $a$ and $b$ are much larger than the thickness of the strip. Determine the force per unit length on the two conductors for a current $I$ flowing through them in opposite directions.
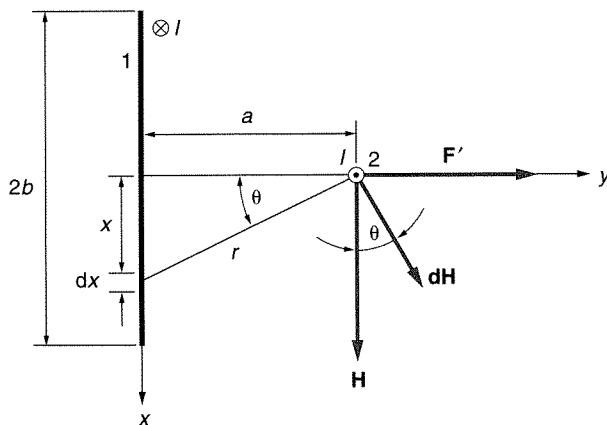


Figure P16.19 Cross section of a two-conductor line
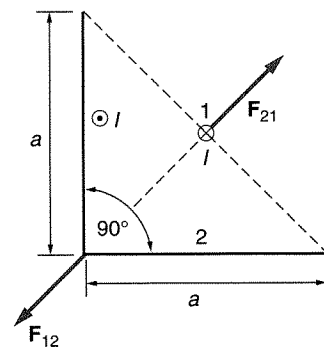
Figure P16.20 Cross section of a two-conductor line

**P16.20.** Determine the force per unit length on the conductors of the line with a cross section as shown in Fig. P16.20. The current in the conductors is $I$ and the medium is air.

*P16.21. Determine the force per unit length on the conductors of the stripline with a cross section as shown in Fig. P16.21. The current in the conductors is $I$, in opposite directions.
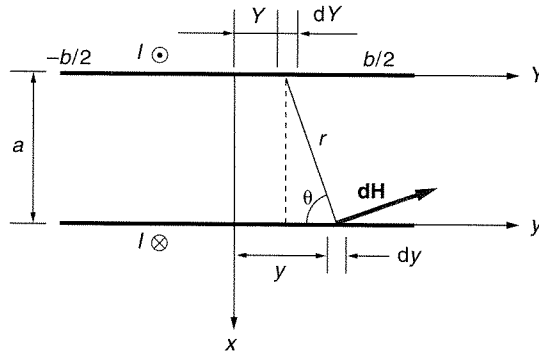


Figure P16.21 Cross section of a stripline

P16.22. A thin two-wire line has conductors of circular cross section of radius $a$ and the distance between their axes $d$ and is short-circuited by a straight conducting bar, as shown in Fig. P16.22. If a current $I$ flows through the line, what is the force on the bar? Assume that the section of the line to the left of the bar is very long, and that the medium is air.
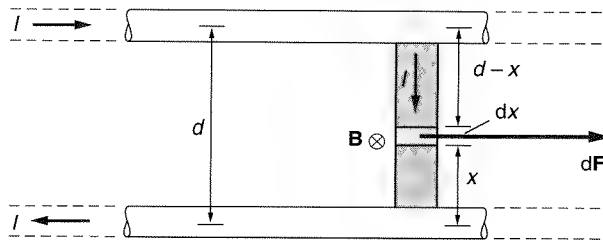


Figure P16.22 Short-circuited two-wire line

P16.23. A long air-filled coaxial cable is short-circuited at its end by a thin conducting plate, as shown in Fig. P16.23. Determine the force on the end plate corresponding to a current of intensity $I$ through the cable.
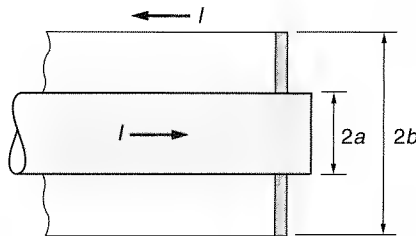


Figure P16.23 Short-circuited coaxial cable

**P16.24.** Determine approximately the force between two parallel coaxial circular loops with currents $I_1$ and $I_2$. The radii of the loops are $a$ and $b$, respectively, with $a \gg b$ and the distance between their centers $z$.

**P16.25.** A metal strip of conductivity $\sigma$ and of small thickness $b$ moves with a uniform velocity $v$ between the round poles of a permanent magnet. The radius of the poles of the magnet is $a$ and the width of the strip is much larger than $a$ (Fig. P16.25). The flux density **B** is very nearly constant over the circle shown hatched and practically zero outside it. Assuming that the induced current density in that circle is given by $\mathbf{J} = \sigma \mathbf{v} \times \mathbf{B}/2$, determine the force on the strip.
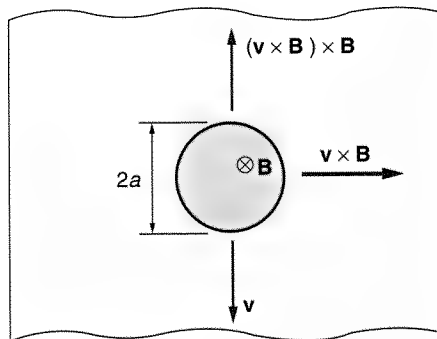


Figure P16.25 A strip moving in a magnetic field

**P16.26.** A thin metal plate is falling between the poles of a permanent magnet (Fig. P16.26) under the action of the gravitational field. The pole radius is $a = 2\,\text{cm}$ and the flux density in the gap is $B = 1\,\text{T}$. Determine approximately the velocity of the plate if its thickness is $b = 0.5\,\text{mm}$, its surface area $S = 100\,\text{cm}^2$, its conductivity $\sigma = 36 \cdot 10^6\,\text{S/m}$ (aluminum), and its mass density $\rho_m = 2.7\,\text{g/cm}^3$ (aluminum).
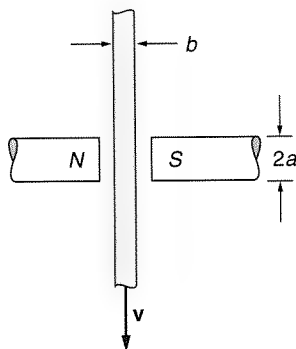


Figure P16.26 A plate falling in a magnetic field

**P16.27.** A metal ring $K$ of negligible resistance is placed above a short cylindrical electromagnet, as shown in Fig. P16.27. Determine qualitatively the time dependence of the

total force on the ring if the current through the electromagnet coil is of the form $i(t) = I_m \cos \omega t$.
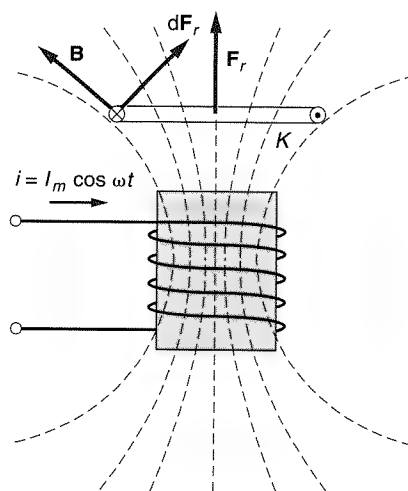


Figure P16.27 A ring in a magnetic field

**P16.28.** A U-shaped glass tube is filled with a paramagnetic liquid of unknown magnetic susceptibility $\chi_m$. A part of the tube inside the dashed square in Fig. P16.28 is exposed to a uniform magnetic field of intensity $H$. Under the influence of magnetic forces, a difference $h$ of the levels of the liquid in the two sections of the tube is observed. Given that the mass density of the liquid is $\rho_m$ and that the medium above the liquid is air, determine $\chi_m$.
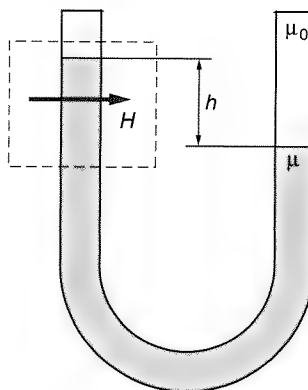


Figure P16.28 A U-shaped tube in a magnetic field

**P16.29.** Plot the scale calibration curve $F_{\text{tot}}(I)$ for the ammeter sketched in Fig. P16.29. $F_{\text{tot}}(I)$ is the total force acting on the iron nail for a given current $I$ in the coil. Given are

$a = 1\,\text{mm}$, $l = 5\,\text{cm}$, $N' = 10$ turns/cm (you need to look up the relative permeability for iron and its mass density).
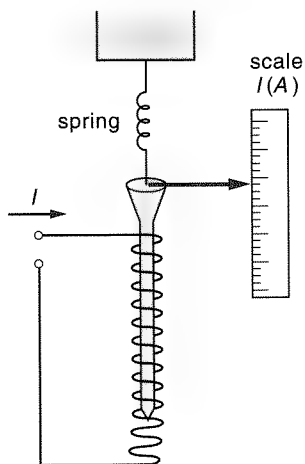


Figure P16.29 Sketch of a simple ammeter

*P16.30. Derive the general expression for pressure of magnetic forces, Eq. (16.22).

# 17

# Some Examples and Applications of Time-Invariant and Slowly Time-Varying Magnetic Fields

## 17.1  Introduction

Magnetic fields are present in many practical applications, as well as in the natural world. For example, we are continuously situated in the relatively strong time-invariant (or extremely slowly variant) magnetic field of the earth; this field may, for example, affect the way your computer monitor works depending on whether you happen to turn it on in the Northern or Southern Hemisphere. We also often find ourselves in magnetic and electric fields existing around high-voltage and high-current power lines, and it is interesting to calculate the order of magnitude of voltages induced in our body. Most electrical home appliances contain devices that use mag-

netic forces and moments of magnetic forces, and our computers, tape recorders, and video recorders use magnetic storage devices. The aim of this chapter is to review some of the more important and interesting applications of magnetic (along with electric) fields.

## 17.2   The Magnetic Field of the Earth

The earth behaves like a large permanent magnet whose magnetic field is similar to the field of a giant current loop with an axis declined 11 degrees with respect to the earth's axis of rotation (Fig. 17.1a). (Geologists believe that the magnetic field is created by the difference in the speed of rotation of the earth's liquid core and its solid mantle.) The planet's geographic North Pole is approximately the *south* magnetic pole (this is why the north pole of the magnetic needle of a compass always points to the north). The magnitude of the earth's magnetic flux density is about 50 $\mu$T at our latitude and about 20% stronger at the poles.

About 90% of the magnetic field measured at the earth's surface is due to the field originating inside the planet. The rest is due to the currents produced by charged particles coming from the sun, and to the magnetism of the rocks in the crust. The region in which the earth's magnetic field can be detected is called the *magnetosphere*. It is not symmetrical, but rather has the shape of a teardrop. This is due to the charged particles streaming from the sun that are deflected by the earth's magnetic field; the earth forms a "shadow" for charges, which has the effect of elongating the magnetosphere.

Because the magnetic field everywhere on the surface of the earth is partly due to magnetization of rocks at that point, magnetometers can be used in geology for de-



(a)                                                             (b)
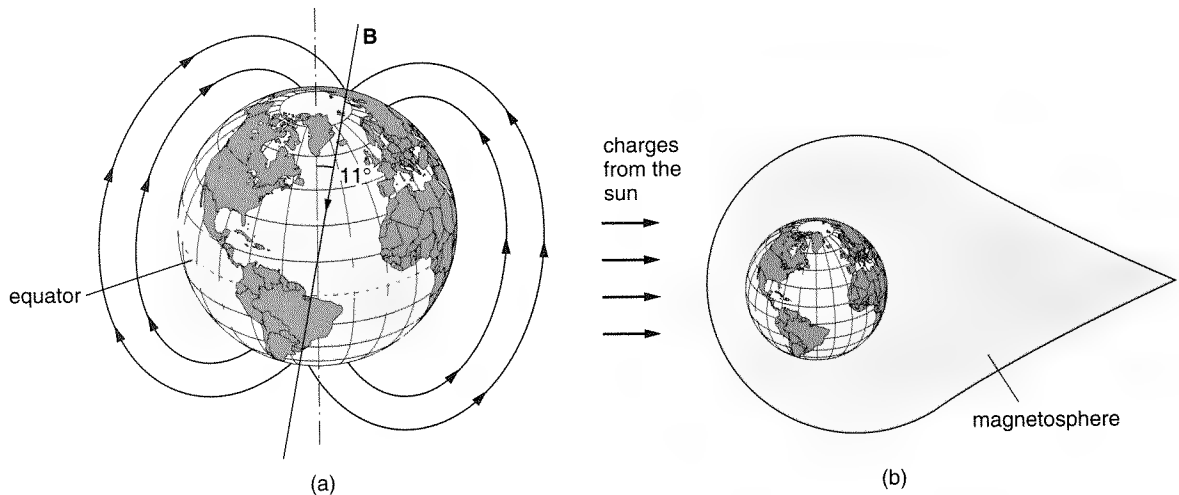
Figure 17.1 (a) The earth produces a dc magnetic field roughly equivalent to the magnetic field of a very large current loop. The plane of the loop is declined with respect to the earth's axis of rotation. (b) The region in which the earth's magnetic field can be detected is called the magnetosphere and is asymmetrical due to the charges emitted from the sun.

tecting different types of ores. Measuring magnetization of rocks also gives us insight into the earth's magnetic history. Rocks become magnetized when they are formed, or when they remelt and recool at some later time. When rocks are heated they lose their magnetization and are remagnetized by the earth's magnetic field as they cool. Therefore, they carry a permanent record of what the earth's magnetic field was like at the time of the rock formation. Measurements of rock magnetization show that the earth's magnetic poles have wandered. Some rocks that formed over short time intervals show fossil magnetic polarities 180 degrees apart, which cannot be explained by a 180-degree rotation of a continent (the time of reversal was too short for this to be possible). The conclusion is that the earth's magnetic field switched polarity, similarly to the field of a loop in which the current changes direction. These field reversals occurred many times during our planet's geological history, and about 10 times in the last 4 million years. Rock magnetization indicates that the polarity does not flip instantly: it first slowly decreases and then increases in the opposite direction.

*Questions and problems:*   Q17.1 to Q17.3

## 17.3   Applications Related to Motion of Charged Particles in Electric and Magnetic Fields

Charged particles in both electric and magnetic fields always move. In many instances one of the fields is of much less influence than the other. For example, we have seen that both an electric and a magnetic field act on moving charges that form an electric current in a conductor, but that the influence of the magnetic field is negligible. There are examples of the other kind, where electric forces exist but are negligible. In some cases, the effects of electric and magnetic fields on a moving particle are of the same order of magnitude and must both be taken into account.

The motion of charged particles in electric and magnetic fields may be in a vacuum (or very rarefied gas), in gases, and in solid or liquid conductors. This brief section is aimed at explaining the principles of motion of charged particles in electric and magnetic fields and at presenting examples of how some engineering applications take advantage of this motion.

We know that the force on a charge $Q$ moving in an electric and a magnetic field with a velocity $\mathbf{v}$ is the Lorentz force, Eq. (12.13), which is repeated here for convenience:

$$\mathbf{F} = Q\mathbf{E} + Q\mathbf{v} \times \mathbf{B}. \tag{17.1}$$

If the electric field can be neglected, we omit the first term of the Lorentz force. If the magnetic field can be neglected, we omit the second term.

If a charge moves in a vacuum, then this force in any instant must be equal in magnitude and opposite in direction to the inertial force. If the mass of the charge is $m$, the equation of motion therefore has the form

$$m\frac{d\mathbf{v}}{dt} = Q\mathbf{E} + Q\mathbf{v} \times \mathbf{B}. \tag{17.2}$$

In this equation, **E** and **B** in general are functions of space coordinates and of time. Except in rare cases, it is impossible to solve such a general equation for the velocity of the charge analytically, but it can always be solved numerically.

If a charge moves in a material (a gas, a liquid, or a solid), collisions influence the (macroscopic) charge motion to a great extent. For example, we have seen that in solid and liquid conductors the motion of free charges is always along the lines of vector **E**. An equation like (17.2) is not valid for average (drift) velocity.

We now discuss a few examples of the motion of charged particles in an electric and a magnetic field.

**Example 17.1—Motion of a charged particle in a uniform electric field.** In Example 11.1 we analyzed the simplest case of motion of a charged particle in a uniform electric field. Let us now consider a more general case, when a charge $Q$ ($Q > 0$) moves in a uniform field with arbitrary initial velocity $\mathbf{v}_0 = \mathbf{v}_{0x} + \mathbf{v}_{0y}$, as in Fig. 17.2a.

The equation of motion (17.2) becomes $m(d\mathbf{v}/dt) = Q\mathbf{E}$. Integrating the scalar $x$ and $y$ components of this equation twice, with respect to the position of the charge as a function of time, we obtain

$$x(t) = \frac{QE}{2m}t^2 + v_{0x}t + x_0 \quad \text{and} \quad y(t) = v_{0y}t + y_0,$$

where $x_0$ and $y_0$ are the initial $x$ and $y$ coordinates of the charge. Consequently, the charge will move along a parabola. This is the same as when we throw a stone at an angle (other than 90 degrees) with respect to the earth's surface.

**Example 17.2—Deflection of an electron stream by a charged capacitor.** Imagine that we shoot an electron between the plates of a charged parallel-plate capacitor, perpendicularly to the electric field. We now know that the electron trajectory will curve toward the positive electrode. So if we put a screen behind the capacitor, with no voltage on the electrodes the
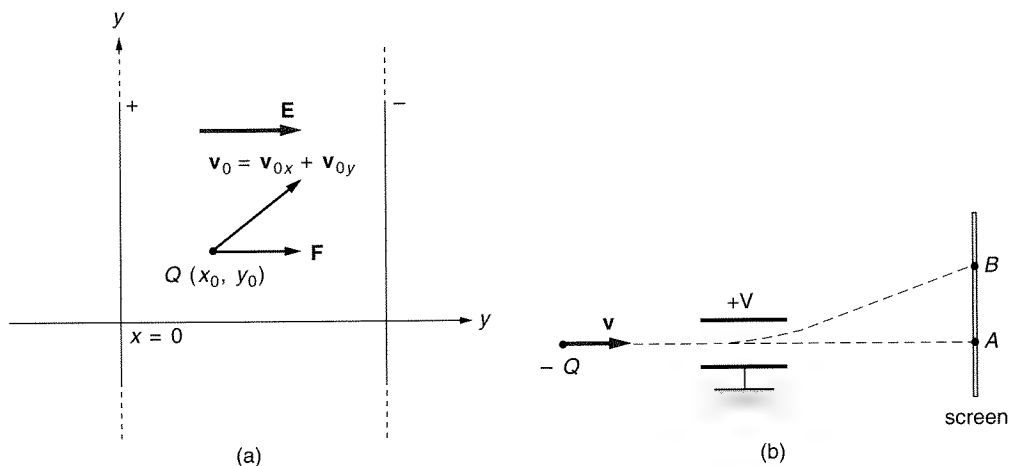


Figure 17.2 (a) Charged particle in a uniform electric field, revisited; (b) deflection of a charged particle by a charged capacitor
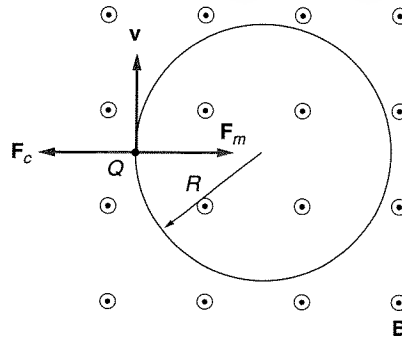
Figure 17.3 Charged particle in a uniform magnetic field

electron hits point $A$ in Fig. 17.2b. When a voltage is applied, the electron is deflected and hits point $B$ on the screen in Fig. 17.2b. Some cathode-ray tubes use this principle for deflecting the electron beam, although the practical deflections are rather small.

**Example 17.3—Motion of a charged particle in a uniform magnetic field.** Consider a charged particle $Q$ ($Q > 0$) moving in a magnetic field of flux density **B** with a velocity **v** normal to the lines of vector **B**, as in Fig. 17.3.

Since the magnetic force on the charge is $\mathbf{F}_m = Q\mathbf{v} \times \mathbf{B}$, it is *always perpendicular to the direction of motion.* This means that a magnetic field cannot change the magnitude of the velocity (i.e., it cannot speed up or slow down charged bodies); it can only change the direction of the charged particle motion. In other words, magnetic forces cannot change the kinetic energy of moving charges.

In the case considered in Fig. 17.3, there is a magnetic force on the charged particle directed as indicated, tending to curve the particle trajectory. Since **v** is normal to **B**, the force magnitude is simply $QvB$. It is opposed by the centrifugal force, $mv^2/R$, where $R$ is the radius of curvature of the trajectory. Therefore, we have

$$QvB = \frac{mv^2}{R},$$

so that the radius of curvature is constant, $R = (mv)/(QB)$. Thus the particle moves in a circle. It makes a full circle in

$$t = T = \frac{2\pi R}{v} = \frac{2\pi m}{QB}.$$

which means that the frequency of rotation of the particle is equal to $f = 1/T = (QB)/(2\pi m)$. Note that $f$ does not depend on $v$. Consequently, all particles that have the same charge and mass make the same number of revolutions per second. This frequency is called the *cyclotron frequency.*

**Example 17.4—The cyclotron.** The cyclotron is a device used for accelerating charged particles. It is sketched in Fig. 17.4. The main part of the cyclotron is a flat metal cylinder, cut along its middle. The two halves of the cylinder are connected to the terminals of an oscillator
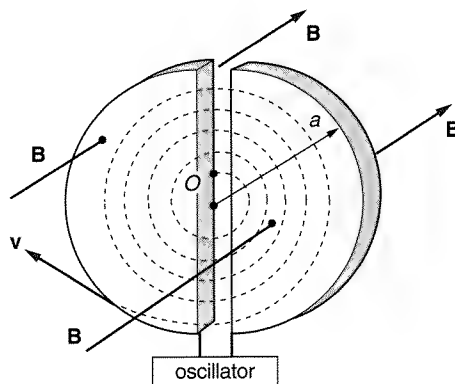
Figure 17.4 A cyclotron

(a source of periodically changing voltage). The whole system is in a uniform magnetic field normal to the bases of the cylinder, and inside the cylinder is a vacuum (i.e., highly rarefied air).

A charged particle from source $O$ finds itself in an electric field that exists between the halves of the cylinder, and it accelerates toward the other half of the cylinder. While outside of the space between the two cylinder halves, the charge finds itself only in a magnetic field, and it circles around with a radius of curvature found as in the preceding example. We saw that the time the charge takes to go around a semicircle does not depend on its velocity. That means that the charge will always take the same amount of time to again reach the gap between the two cylinders. If the electric field variation in this region is adjusted in such a way that the charge is always accelerated, the charge will circle around larger and larger circles, with increasingly larger velocity, until it finally shoots out of the cyclotron. The velocity of the charge when it gets out of the cyclotron is $v = (QBa)/m$. This equation is valid only for velocities not close to the speed of light. If this is not the case, the relativistic effects increase the mass, i.e., the mass is not constant.

As a numerical example, for $B = 1\,\mathrm{T}$, $Q = e$, $a = 0.5\,\mathrm{m}$, $m = 1.672 \cdot 10^{-27}\,\mathrm{kg}$ (a proton), we get $v = 47.9 \cdot 10^6\,\mathrm{m/s}$.

**Example 17.5—Cathode-ray tube.** Cathode-ray tubes (CRTs), used in some TVs and computer monitors, have controlled electron beams that show traces on a screen. One system for deflecting electron streams in CRTs is sketched in Fig. 17.2b. Basically, there are two mutually orthogonal parallel-plate capacitors, which can deflect the stream in two orthogonal directions. In this way the electron stream can hit any point of the screen, and precisely where it hits is controlled by appropriate voltages between the electrodes of the two capacitors.

We have already mentioned that the electric field can deflect electron streams only by relatively small distances. When a large deflection is required, as in television receivers, a magnetic field is used, as sketched in Fig. 17.5. The design of the magnetic deflecting system (a coil of a specific geometry and with many turns of wire) is rather complicated and is usually done experimentally. Part of the experimental adjustment is due to the effect of the earth's magnetic field on charged particles at any point on the planet.

If we think of the earth as of an equivalent current loop, as described in section 17.2, the horizontal component of the magnetic flux density vector is oriented along the north-south direction, and the vertical component is oriented downward in the Northern Hemisphere and
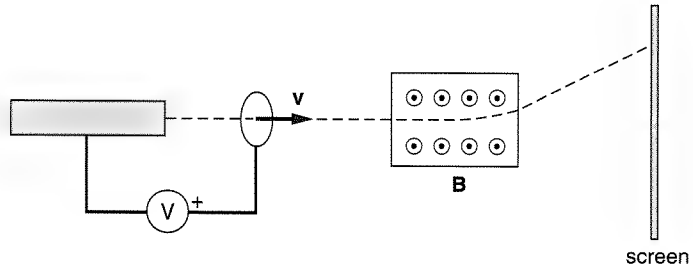
Figure 17.5 A system for forming a stream of electrons and then deflecting it with a magnetic field

upward in the Southern Hemisphere. Therefore, CRTs that use magnetic field deflection have to be tuned to take this external field into account. It is likely that your computer monitor (if a CRT) will not work exactly the same way if you turn it sideways (it might slightly change colors or shift the beam by a couple of millimeters), or if you use it in the other hemisphere of the globe.

Example 17.6—The Hall effect. In 1879, Edwin Hall thought of a clever way of determining the sign of free charges in conductors. A ribbon made of the conductor we are interested in has a width $d$ and is in a uniform magnetic field of flux density $\mathbf{B}$ perpendicular to the ribbon (Fig. 17.6). A current of density $\mathbf{J}$ flows through the ribbon. The free charges can in principle be positive, as in Fig. 17.6a, or negative, as in Fig. 17.6b. The charges that form the current are moving in a magnetic field, and therefore a magnetic force $\mathbf{F} = Q\mathbf{v} \times \mathbf{B}$ is acting on them. Due to this force, positive charges accumulate on one side of the ribbon, and negative ones on the other side. These accumulated charges produce an electric field $E_H$. This electric field, in turn, acts on the free charges with a force that is in the opposite direction to the magnetic force. The charges will stop accumulating when the electric force is equal in magnitude
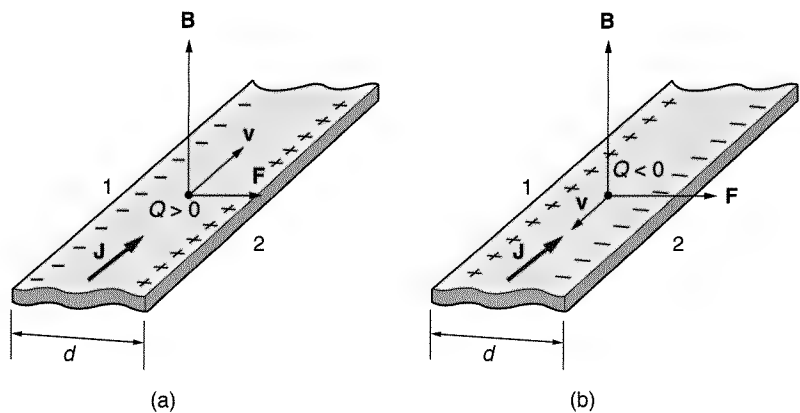


Figure 17.6 The Hall effect in case of (a) positive free charge carriers, and (b) negative free charge carriers

to the magnetic force acting on each of the charges. So in the steady state,

$$QvB = QE_H, \quad \text{or} \quad E_H = vB.$$

Between the left and right edge of the ribbon, we can measure a voltage equal to

$$|V_{12}| = E_H d = vBd.$$

In the case shown in Fig. 17.6a this voltage is negative, and in Fig. 17.6b it is positive. So the sign of the voltage tells us the sign of free charge carriers, and a voltmeter can be used to determine this sign.

Since $J = NQv$, where $N$ is the number of free charges per unit volume, we can write

$$|V_{12}| = \frac{Jd}{NQ} B.$$

Thus if we determine the coefficient $Jd/NQ$ for either ribbon sketched in Fig. 17.6 (which is usually done experimentally), by measuring $V_{12}$ we can measure $\mathbf{B}$. This ribbon has four terminals—two for the connection to a source producing current in the ribbon, and two for the measurement of voltage across it. Such a ribbon is called a *Hall element.*

For single valence metals, e.g., copper, if we assume that there is one free electron per atom, the charge concentration is given by

$$N = \frac{N_A \rho_m}{M},$$

where $N_A$ is Avogadro's number ($6.02 \cdot 10^{23}$ atoms/mole), $\rho_m$ is the mass density of the metal, and $M$ is the atomic mass.

*Questions and problems:*  Q17.4 to Q17.7, P17.1 to P17.5

## 17.4  Magnetic Storage

Magnetic materials have been used for storing data since the very first computers. The first computer memories consisted of small toroidal ferromagnetic cores arranged in two-dimensional arrays, in which digital information was stored in the form of magnetization. These memories are bulky and slow, as can be concluded from their description in Example 17.7. Today's memories are essentially electrostatic: capacitances inside transistors are used for storing bits of information in the form of charges.

The hard disk in every computer is also a magnetic memory. We can write to the disk by magnetizing a small piece of the disk surface, and we can read from the disk by inducing a voltage in a small loop that is moving in close proximity to the magnetized disk surface element. As technology has improved, the amount of information that can be stored on a standard-size hard disk has grown rapidly. Between 1995 and 1997 the standard capacity of hard disks on new personal computers shot from a few hundred megabytes to more than 2 gigabytes. The development is in the

direction not only of increasing disk capacity but also of increasing speed (or reducing the access time). As we will see in Example 17.8, these two requirements compete with each other, and the engineering solution, as is usually the case, needs to be a compromise.

**Example 17.7—History: magnetic core memories.** Magnetic core memories were used in computers around 1970 but are now completely obsolete. The principle of their operation, however, is clever.

A magnetic core memory uses the hysteresis properties of ferromagnetics. One "bit" of the memory is a small ferromagnetic torus, shown in Fig. 17.7a. Two wires, in circuits 1 and 2, pass through the torus. Circuit 1 is used for writing and reading, and circuit 2 is used only for reading. To write, a positive ("1") or negative ("0") current pulse is passed through circuit 1 in the figure. When a positive current pulse is sent through the circuit, the core is magnetized to the point labeled $B_r$ on the hysteresis curve in Fig. 17.7b. When a negative current pulse is passed through the circuit, the core is magnetized to the point labeled $-B_r$. So the point $B_r$ corresponds to a "1," and $-B_r$ to a "0."



(a)

(b)

(c)
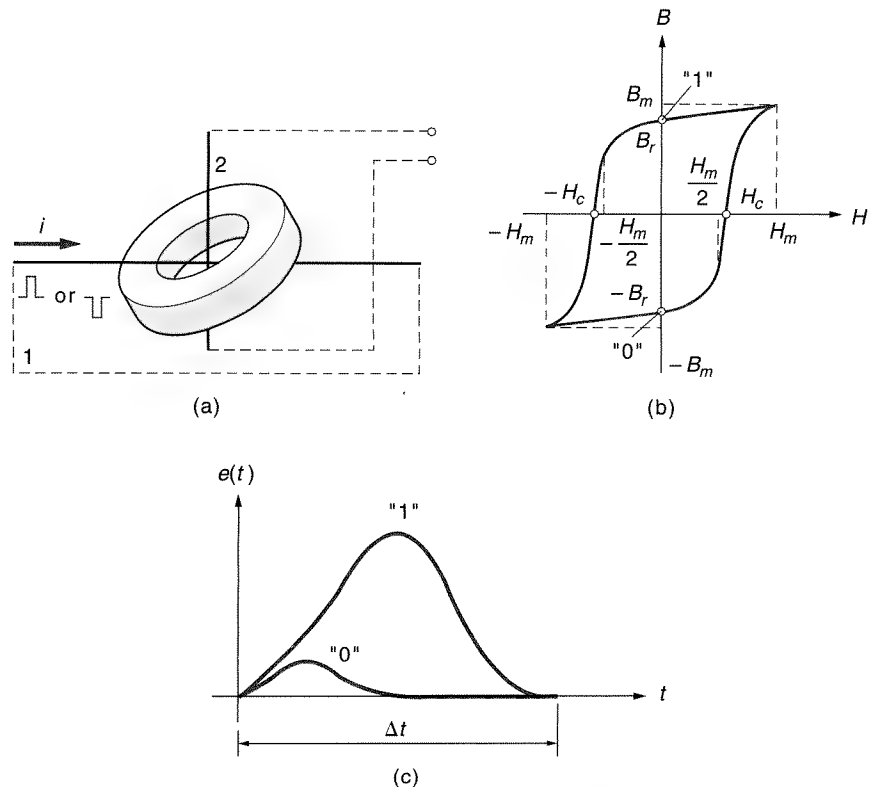
Figure 17.7 (a) One bit of a magnetic core memory. (b) Hysteresis loop of the core. (c) The induced emf pulses produced in circuit 2 while reading out the binary value written in the core
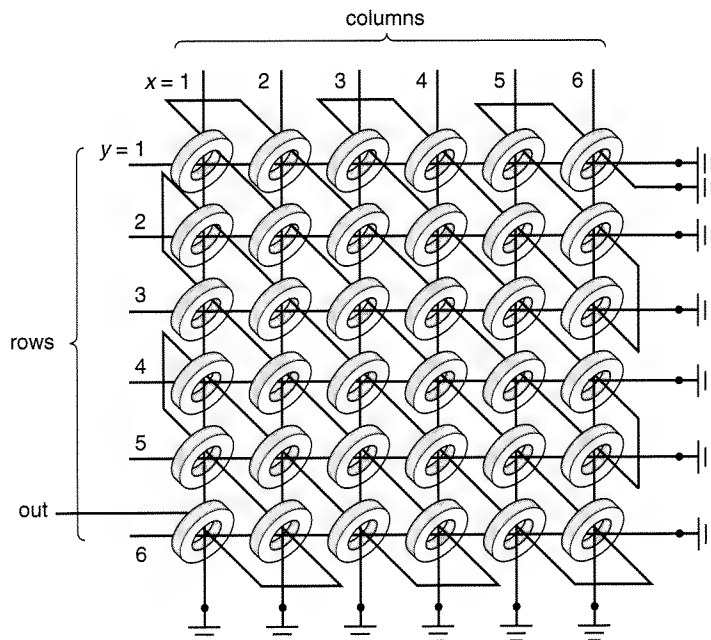
Figure 17.8 A sketch of a magnetic memory

How is the reading performed? A negative current pulse is passed through circuit 1. If the core is at a "1," the negative current pulse will bring the operating point to $-B_m$ (the negative tip of the loop), and after the pulse is over the point will move to $-B_r$ on the hysteresis loop. If the core is at a "0," the negative pulse will make the point go to the negative tip of the loop, and then return to $-B_r$.

While this is done, an emf is induced in circuit 2, resulting in one of the two possible readings shown in Fig. 17.7c. These two "pulses" correspond to a "1" and a "0." The speed at which this is done is about $\Delta t = 0.5 - 5\,\mu\text{s}$. The dimensions of the torus are small: the outer diameter is 0.55 to 2 mm, the inner diameter is 0.3 to 1.3 mm, and the thickness is 0.12 to 0.56 mm.

Elements of an entire memory are arranged in matrices, as shown in Fig. 17.8. Two wires pass through each torus, as in Fig. 17.7a. The current passing through each row or column is only half the current needed to saturate the torus, so both the row and column of a specific core need to be addressed.

**Example 17.8—Computer hard disks.** The hard disk in every computer has information written to and read from it. We will describe how both processes work in modern hard disks and discuss some of the engineering parameters important for hard disk design. The hard disk itself is coated with a thin coating of ferromagnetic material such as $Fe_2O_3$. The disk is organized in sectors and tracks, as shown in Fig. 17.9.

The device that writes data to the disk and reads data from it is called a *magnetic head*. Magnetic heads are made in many different shapes, but all operate according to the same principle. We will describe the read-write process for a simplified head, shown in Fig. 17.10. It is a magnetic circuit with a gap. The gap is in close proximity to the tracks, so there is some leakage flux between the head and the ferromagnetic track.
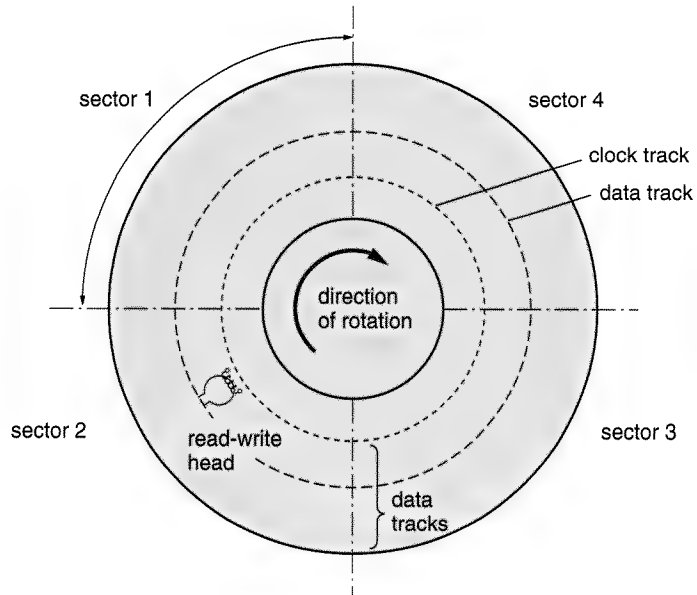
Figure 17.9 Hard disk tracks
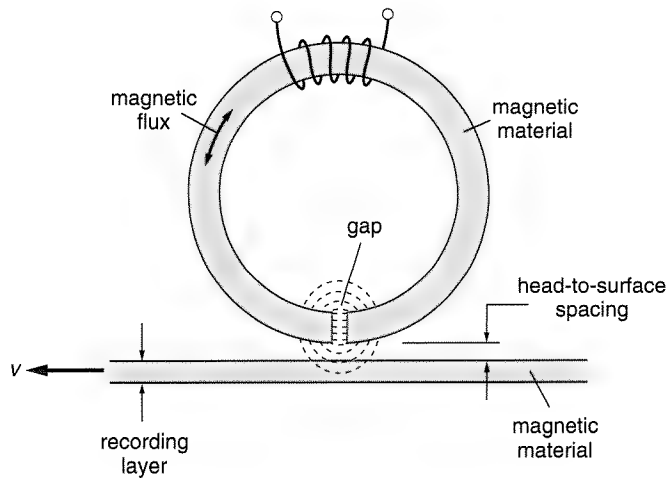


Figure 17.10 Magnetic head

In the "write" process, a current flows through the winding of the magnetic head, thus creating a fringing magnetic field in the gap. The gap is as small as $5\,\mu$m. As the head moves along the track (usually the track rotates), the fringing field magnetizes a small part of the track, creating a south and a north pole in the direction of rotation. These small magnets are about $5\,\mu$m long by $25\,\mu$m wide. A critical design parameter is the height of the head above the track: the head cannot hit the track and get smashed, but it also needs to be as close as possible to maximize the leakage flux that magnetizes the track. Typically, the surface of the track is flat
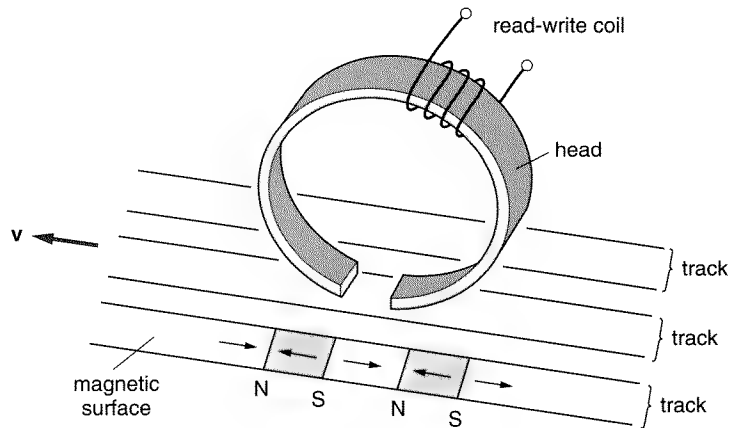
Figure 17.11 The magnetic head aerodynamically flies over the disk surface at a distance of only about 1 micrometer above it, following the surface profile.

to within several micrometers, and the head follows the surface profile at a distance of about 1 micrometer or less above it. This is possible because the head aerodynamically flies above the disk surface, as shown in Fig. 17.11. The current in the head windings should be strong enough to saturate the ferromagnetic track. If the track is saturated, the voltage signal during readout is maximized.

In the "read" process, there is no current in the windings of the magnetic head. The residual magnetization of the magnetic head should be as small as possible, so that the head is demagnetized when the current is turned off in readout. Now the flux from the magnetized track induces a voltage between the winding open ends while the head is moving with respect to the track (according to Faraday's law). Since the largest changes in **B** occur when the magnetic field changes direction, i.e., between two tiny magnets along the track, the output voltage has a waveform consisting of positive and negative pulses, as shown in Fig. 17.12b. The volt-
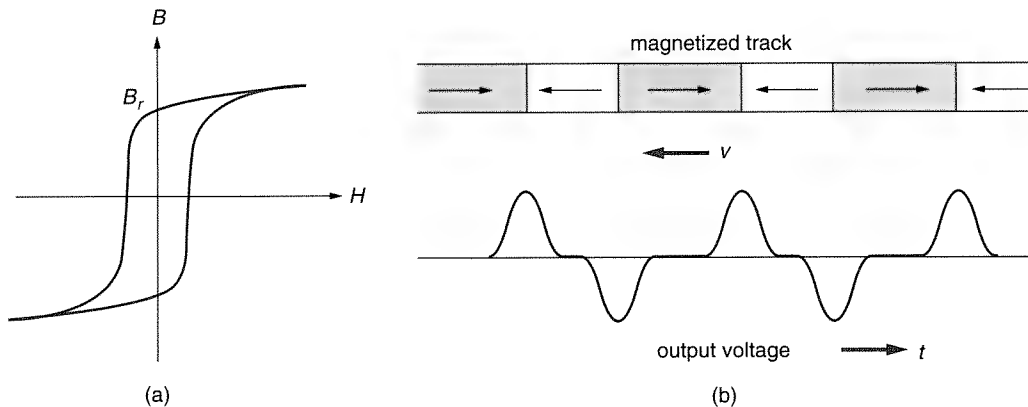


Figure 17.12 (a) Hysteresis curve of the track material. $B_r$ is the remanent magnetic flux density and should be large for good readout. (b) A typical voltage signal read from the disk.

age is proportional to the remanent magnetic flux density, $B_r$, of the ferromagnetic hysteresis curve in Fig. 17.12a.

The capacity of data storage is given by the information density per unit area of storage surface. The storage density per unit surface area is the product of the storage density per unit track length, times the track density per unit distance normal to the direction of relative motion. An increase in track density reduces the sharpness in magnetic field discontinuity, thus reducing the readout voltage. Note that magnetic disks are inherently binary storage devices and that the frequency of the voltage pulses in readout is doubled compared to the number of actual segments along the track.

*Questions and problems:*   Q17.8 and Q17.9, P17.6 and P17.7

## 17.5  Transformers

A transformer is a magnetic circuit with (usually) two windings, "primary" and "secondary," on a common ferromagnetic core (Fig. 17.13a). When an ac voltage is applied to the primary coil, the magnetic flux through the core is the same at the secondary and induces a voltage at the open ends of the secondary winding. Ampère's law for this circuit can be written as

$$N_1 i_1 - N_2 i_2 = Hl,$$

where $N_1$ and $N_2$ are the numbers of the primary and secondary turns, $i_1$ and $i_2$ are the currents in the primary and secondary coils when a generator is connected to the primary and a load to the secondary, $H$ is the magnetic field in the core, and $l$ is the effective length of the core. Since $H = B/\mu$ and, for an *ideal* core, $\mu \to \infty$, both $B$ and $H$ in the ideal core are zero (otherwise the magnetic energy in the core would be infinite). Therefore for an ideal transformer we have

$$\frac{i_1}{i_2} = \frac{N_2}{N_1}. \tag{17.3}$$
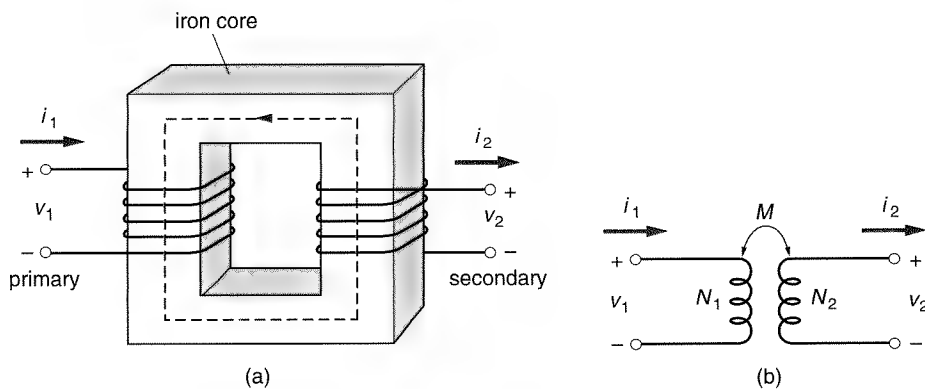


Figure 17.13 (a) A transformer, and (b) the equivalent circuit of an ideal transformer

This is the relationship between the primary and secondary currents in an *ideal transformer*. For good ferromagnetic cores, the permeability is high enough that this is a good approximation.

From the definition of magnetic flux, we know that the flux through the core is proportional to the number of turns in the primary. From Faraday's law, we know that the induced emf in the secondary is proportional to the number of times the magnetic flux in the core passes through the surface of the secondary winding, that is, to $N_2$. Therefore, we can write the following for the voltages across the primary and secondary windings:

$$\frac{v_1}{v_2} = \frac{N_1}{N_2}. \tag{17.4}$$

From our discussion of mutual inductance in Chapter 15, we see that the equivalent circuit of an ideal transformer is just a mutual inductance, as shown in Fig. 17.13b.

Assume that the secondary winding of an ideal transformer is connected to a resistor of resistance $R_2$. What is the resistance seen from the primary terminals? From Eqs. (17.3) and (17.4) we obtain

$$R_1 = \frac{v_1}{i_1} = \frac{v_2 N_1/N_2}{i_2 N_2/N_1} = R_2 \left(\frac{N_1}{N_2}\right)^2. \tag{17.5}$$

From this discussion, we see that the transformer's name is appropriate: it transforms the values of the voltage, current, and resistance between the primary and secondary windings. The transformation ratio is dictated by the ratio of the number of turns. In an ideal transformer there are no losses, so all of the power delivered to the primary can be delivered to a load connected to the secondary.

In a realistic transformer there are several loss mechanisms: resistance in the wire of the windings, and eddy current losses and hysteresis losses in the ferromagnetic core. To minimize resistive losses in sometimes very long wires used for a large number of turns, a good metal such as copper is chosen. Eddy current losses are minimized by laminating the core, as discussed in Example 14.6, and hysteresis losses were discussed in Example 16.3. These losses, as well as the inductance of the windings, result in a realistic equivalent circuit for a transformer shown in Fig. 17.14. For
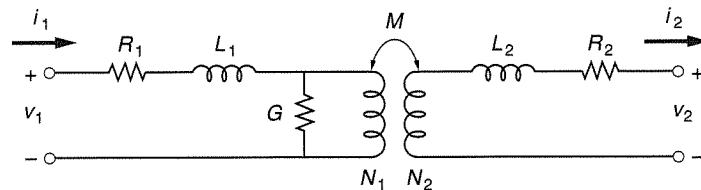


Figure 17.14 Equivalent circuit of a realistic transformer

high-frequency transformers, or in cases where transients are important, the capacitance between the winding turns also needs to be taken into account.

*Questions and problems:*    Q17.10 and Q17.11, P17.8

## 17.6    Synchronous and Asynchronous (Induction) Electric Motors

Electric motors serve to continuously transform electric energy into mechanical energy. There are several types of electric motors, and we will briefly describe two types that use the concept of rotating magnetic fields.

Imagine we have a U-shaped magnet that rotates with an angular velocity $\omega$, as in Fig. 17.15. The magnetic field will rotate with the magnet; thus it is known as the *rotating magnetic field*. (We will see that a rotating magnetic field can be obtained with appropriate sinusoidal currents in *stationary coils*.) Let a small magnet, e.g., a compass needle, be situated in this field, with the axis of rotation the same as that of the U-shaped magnet. A magnetic torque will act on the small magnet. If the small magnet is stationary, and $\omega$ is large, there will be a torque on the small magnet that tends to rotate it in one and then in the other direction, so that it will only oscillate. However, if the small magnet is brought to rotate with the angular velocity $\omega$, the rotating magnetic field will act on it by a continuous torque in one direction, and the small magnet will rotate *in synchronism* with the magnetic field, even if it has to overcome a small friction (or a load). If the rotating magnetic field is obtained with currents in stationary coils, the same will happen, and we will have a simple *synchronous motor*. The name comes from the fact that the motor can rotate only in synchronism with the rotation of the magnetic field.

If instead of the small magnet we have in the rotating magnetic field a short-circuited wire loop, as in Fig. 17.16, a current will be induced in the loop because the magnetic flux through the loop is varying in time. According to Lentz's law, the actual direction of the induced current will be as indicated in the figure. It is seen that there will be a torque on the loop, tending to rotate it with the field. If the rotating field is produced by currents in stationary coils, we obtain a simple *induction motor*.
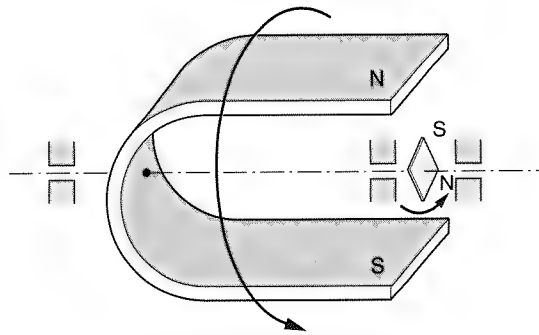


Figure 17.15 A small magnet in the rotating magnetic field of a rotating magnet
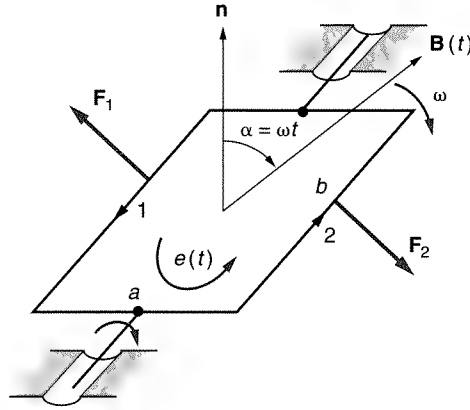
Figure 17.16 A short-circuited wire loop in a rotating magnetic field

Because there will be a time-average torque on the loop for any angular velocity of the loop rotation, whether or not it rotates in synchronism with the field, it is also known as the *asynchronous* (i.e., not synchronous) *motor*. Its rotating part, or *rotor*, is usually made in the form of a number of short-circuited loops at an angle, similar to a cage (the *squirrel-cage rotor*). The short-circuited loops are fixed in grooves in a ferromagnetic rotor core.

For large amounts of power, the rotating magnetic field is obtained directly from a three-phase current system. Let us examine a simpler way of obtaining a rotating magnetic field using two currents of equal amplitude that are 90 degrees out of phase (Fig. 17.17), a method used for low-power synchronous and asynchronous motors.

If the currents in the two coils in Fig. 17.17 are of equal magnitude and shifted in phase by 90 degrees, so are the magnetic flux densities they produce. Therefore (see Fig. 17.17),

$$B_x(t) = B_m \cos \omega t, \qquad \text{and} \qquad B_y(t) = B_m \sin \omega t.$$

The total magnetic flux density has a magnitude of

$$B_{\text{total}}(t) = \sqrt{B_x^2(t) + B_y^2(t)} = B_m,$$

which means that it has a constant magnitude. The vector **B** is rotating, however, since

$$\tan \alpha(t) = \frac{B_y(t)}{B_x(t)} = \tan \omega t,$$

so that

$$\alpha(t) = \omega t,$$
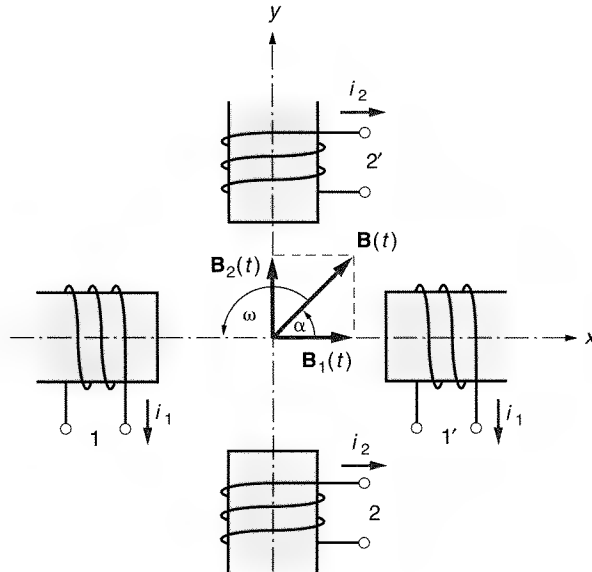
**Figure 17.17** A rotating magnetic field produced with two perpendicular coils with sinusoidal currents shifted in phase by 90 degrees

which means that, indeed, the vector **B** rotates with a constant angular velocity $\omega$. We have thus obtained a rotating magnetic field with sinusoidal currents in two stationary coils.

Three-phase motors and generators were invented by Nikola Tesla, a Serbian immigrant who came to America with 4 cents in his pocket. In 1891, he filed about 50 patents related to different kinds of ac generators and motors, but he had to fight Thomas Edison's promotion of dc power. Eventually George Westinghouse, who supported Tesla's inventions, won the battle and the first ac power plant was built on Niagara Falls. In 1891, the same year Tesla filed the patents that provoked strong reactions and resistance in the scientific community, mines in Telluride, Colorado, had already installed polyphase motors and generators based on his patents.

*Questions and problems:*   Q17.12 and Q17.13, P17.9

## 17.7   Rough Calculation of the Effect of Power Lines on the Human Body

We often hear that the electric and magnetic fields "radiated" from power lines may be harming our bodies. We are now ready to do some rough calculations of the voltages induced in the body. We will do the calculations on the example of a human head (which is the most important and most sensitive part of our body). We will assume that our head is a sphere with a radius of 10 cm, and made mostly of salty water. In this example, let the power lines be as close as 20 m to a human, and let them carry an
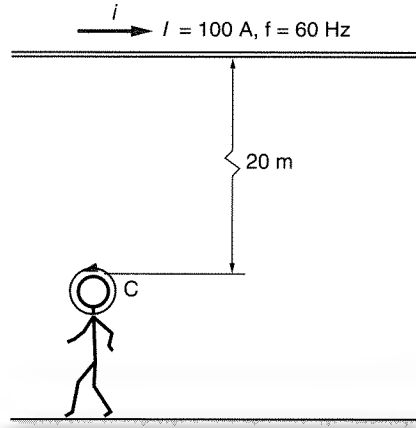
Figure 17.18 The effect of a power line on the human body

unbalanced current of 100 A, as in Fig. 17.18. (The total current in a balanced power line is zero.)

We need to consider two effects: the induced voltage from the magnetic field, since the current in the line is sinusoidally varying in time at 60 Hz, and the voltage due to the electric field of the wire.

First, the magnetic flux density 20 m from a wire carrying 100 A is equal to $B = \mu_0 I / 2\pi r = 1\,\mu\text{T}$. For comparison, the earth's average dc magnetic field is $50\,\mu\text{T}$ on the earth's surface. (This dc field induces currents in our bodies only if we move, and we are probably adapted to this small effect.) The induced electric field around our head (which is a conductor) can be calculated from Faraday's law:

$$\oint_{\text{head perimeter}} \mathbf{E}_{\text{ind}} \cdot \mathbf{dl} = -\frac{\partial}{\partial t} \int_{\text{head cross section}} \mathbf{B} \cdot \mathbf{dS}.$$

The left-hand side of the equation is approximately $2\pi a E_{\text{ind}}$, and the right-hand side equals $-a^2 \pi\, \partial B / \partial t$. In complex notation we thus have

$$2\pi a E_{\text{ind}} = -j\omega B\pi a^2, \qquad \text{hence} \qquad |E|_{\text{ind}} = \omega B a / 2. \tag{17.6}$$

The rms value of the voltage due to this induced electric field across a single cell in our head (which is about $10\,\mu\text{m}$ wide) is $V_{\text{cell}} \simeq 33\,\text{pV}$ at 60 Hz. This is very small; for comparison, the normal neural impulses that pass through the cells are spikes on the order of 100 mV in amplitude, and they last about a millisecond, with a frequency between 1 and 100 Hz.

Let us now consider the electric field effect. The approximate value for the electric field around power lines depends on the power line voltage rating and the distance of the point from the line, but a resonable value would be $E_0 = 1\,\text{kV/m}$. Our cells are made of essentially salty water, which has a resistivity of about $1\,\Omega \cdot \text{m}$. To find the voltage across an individual cell we reason as follows.

The sphere approximating the head is conducting. It is situated in an approximately uniform electric field. Therefore, surface charges are induced on its surface as determined in Example 11.3, Eq. (11.9):

$$\sigma(\theta) = 3\epsilon_0 E_0 \cos\theta,$$

where $\theta$ is the angle between the radius to the cell considered (a point on the sphere surface) and the direction of vector $\mathbf{E}_0$. However, this charge is not time-constant, as in Example 11.3, but rather time-varying, since $E_0$ is time-varying. Consequently, there is a time-varying current inside the sphere, which can be determined approximately in the following manner.

The total charge on one hemisphere is given by

$$Q = \int_0^{\pi/2} \sigma(\theta) 2\pi a \sin\theta \, a \, d\theta = \int_0^{\pi/2} 3\epsilon_0 E \cos\theta \, 2\pi a \, \sin\theta \, a \, d\theta = 3\pi\epsilon_0 a^2 E_0. \quad (17.7)$$

For $E_0 = 1\,\text{kV/m}$, this amounts to $Q = 835\,\text{pC}$.

If the charge is time-varying, there is a time-varying current in the sphere obtained as $i(t) = dQ(t)/dt$. For a sinusoidal field of frequency $f = 60\,\text{Hz}$, the rms value of the current is

$$I = \omega Q = 2\pi f Q = 0.315\,\mu\text{A}.$$

The current density inside the sphere, in the equatorial plane of the sphere and normal to $\mathbf{E}_0$, is hence

$$J = \frac{I}{a^2\pi} \simeq 10\,\mu\text{A/m}^2,$$

so that the electric field inside the sphere is not zero, but has a rms value $E = \rho J = 10\mu\text{V/m}$. So the voltage across a cell equals about $10\,\mu\text{V/m} \times 10\,\mu\text{m} = 100\,\text{pV}$. This is somewhat larger than the voltage due to the time-varying magnetic field, but it is probably still negligible with respect to 100-mV voltage spikes due to normal neural impulses.

In conclusion, voltages induced in our body when we are close to power lines are much smaller than the normal electric impulses flowing through our nerve cells. Nevertheless, it is hard to say with absolute certainty that these orders-of-magnitude lower voltages do not have any effect on us, because biological systems are often at a very unstable equilibrium.

*Questions and problems:*   Q17.14

# QUESTIONS

**Q17.1.** Where is the earth's south magnetic pole?

**Q17.2.** What is the order of magnitude of the earth's magnetic flux density?

**Q17.3.** Approximately how fast would you need to spin around your axis in the magnetic field of the earth to induce 1 mV around the contour of your body?

**Q17.4.** Turn your computer monitor sideways or upside down while it is on (preferably with some brightly colored pattern on it). Do you notice changes in the screen? If yes, what and why?

**Q17.5.** What do you expect to happen if a magnet is placed close to a monitor? If you have a small magnet, perform the experiment (note that the effect might remain after you remove the magnet, but it is not permanent). Explain.

**Q17.6.** Explain how the Hall effect can be used to measure the magnetic flux density.

**Q17.7.** Explain how the Hall effect can be used to determine whether a semiconductor is *p*- or *n*-doped.

**Q17.8.** What magnetic material properties are chosen for the tracks and heads in a hard disk?

**Q17.9.** Sketch and explain the time-domain waveform of the induced emf (or current) in the magnetic head coil in "read" mode as it passes over a piece of information recorded on a computer disk as "110." (Assume that a "1" is a small magnet along the track with a N-S orientation from left to right, and a "0" is in the opposite direction.)

**Q17.10.** Write Ampère's law for an ideal transformer, and derive the voltage, current, and impedance (resistance) transformation ratio. The number of turns in the primary and secondary are $N_1$ and $N_2$.

**Q17.11.** What are the loss mechanisms in a real transformer, and how does each of the contributors to loss depend on frequency?

**Q17.12.** Explain how a synchronous motor works.

**Q17.13.** How is an asynchronous motor different from the synchronous type?

**Q17.14.** Describe the two mechanisms by which ac currents can affect our body. Use formulas in your description.

# PROBLEMS

**P17.1.** What is the minimum magnitude of a magnetic flux density vector that will produce the same magnetic force on an electron moving at $100\,\text{m/s}$ that a 10-kV/cm electric field produces?

**P17.2.** Calculate the velocity of an electron in a 10-kV CRT. The electric field is used to accelerate the electrons, and the magnetic field to deflect them.

**P17.3.** How large is the magnetic flux density vector needed for a 20-cm deflection in the CRT in problem P17.2, if the length of the tube is 25 cm?

**P17.4.** A thin conductive ribbon is placed perpendicularly to the field lines of a uniform **B** field. When the current is flowing in the direction shown in Fig. 17.6a, there is a measured negative voltage $V_{12}$ between the two edges of the ribbon. Are the free charges in the conductive ribbon positive or negative?

**P17.5.** What is the voltage $V_{12}$ equal to in problem P17.4 if $B = 0.8\,\text{T}$, the ribbon thickness $t = 0.5\,\text{mm}$, $I = 0.8\,\text{A}$, and the concentration of free carriers in the ribbon is $N = 8 \cdot 10^{28}\,\text{m}^{-3}$?

**P17.6.** The magnetic head in Figure 17.10 is in write mode. Calculate the magnitude of the current $i$ in the winding that would be needed to produce a $B_0 = 1\,\mu\text{T}$ field in the gap. There are $N = 5$ turns on the core, the core can approximately be considered as linear, of relative permeability of $\mu_r = 1000$, the gap is $L_0 = 20\,\mu\text{m}$ wide, the cross-sectional

area of the core is $S = 10^{-9} m^2$, and the mean radius of the core is $r = 0.1$ mm. Assume that the fringing field in the gap makes the gap cross-sectional area effectively 10% larger than that of the core.

**P17.7.** The head and the tracks in magnetic hard disks are made of different magnetic materials because they perform different functions. Sketch and explain the preferred hysteresis curves for the two materials, indicating the differences. Which has higher loss in the ac regime?

**P17.8.** A CRT needs 10 kV to produce an electric field for electron acceleration. Design a wall-plug transformer to convert from 110 V in the U.S. and Canada or 220 V in Europe and Asia. Assume you have a core made of a magnetic material that has a very high permeability.

**P17.9.** Assume that in Fig. 17.17 you have three instead of two coils. The axes of the coils are now at 60 degrees, not 90 degrees, with respect to each other. What is the relative phasing of three sinusoidal currents in the coils that will give a rotating magnetic field, as described in section 17.6 for the case of two currents? Plot the current waveforms as a function of time.

# *18*

# *Transmission Lines*

## 18.1  Introduction

We have by now learned what capacitors, resistors, and inductors are from the standpoint of electromagnetic field theory. In circuit theory, we usually assume that these elements are lumped (pointlike) and that they are interconnected by means of wire conductors. The current along a wire conductor is assumed to be the same at all points.

Transmission lines consist most frequently of two conductors (some have more, e.g., a three-phase power line). Examples are a coaxial line, a two-wire line, and a stripline. Transmission lines are rare electromagnetic systems that can also be analyzed by circuit-theory tools, although we need electromagnetic theory for determining the transmission-line parameters (i.e., the circuit elements).

Consider a very long section of a transmission line, such as a coaxial line, with perfect conductors and an imperfect dielectric. Let a dc generator of voltage $V$ be connected at one end of the line and a resistor of resistance $R$ at the other. Is there a current along the line? The answer is, of course, yes. However, because there are stray currents through the imperfect dielectric from one line conductor to the other, the current intensity along the two line conductors is not constant. The largest current will be at the generator end, as at that point all the stray currents add up. At the other end of the line, the current intensity through the conductors is the smallest, equal to $V/R$, as sketched in Fig. 18.1a. Note that if the conductors are perfect, the current intensity in the resistor does not depend on the stray currents.

320
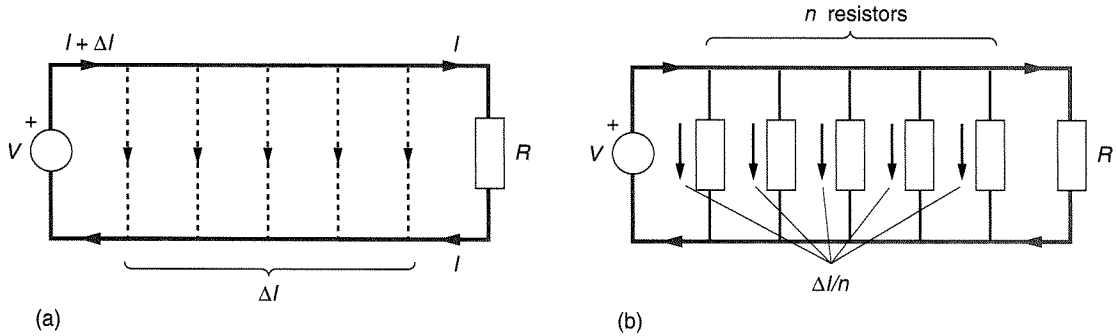
Figure 18.1 (a) A section of a transmission line with perfect conductors and an imperfect dielectric shows stray currents. (b) A ladderlike circuit-theory approximation of the line in (a).
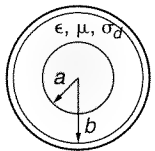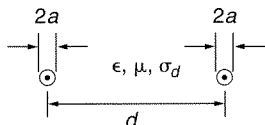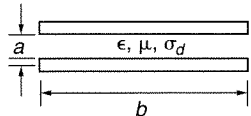
The conclusion that currents at the generator and load ends are different does not fit into the circuit-theory postulate that the current is the same all along a wire that connects circuit elements. Is it possible nevertheless to use circuit theory to analyze this simple circuit? We can subdivide the line section into short segments and represent it as a ladderlike structure with appropriate resistors connected between the conductors of these short line elements, as in Fig. 18.1b. The accuracy of this approximation will increase with the number of segments. For exact representation we need an infinite number of infinitely small segments, but a large number of segments should also give us an accurate result.

If instead of a dc generator we connect an ac generator, the same effect occurs even if the line dielectric is perfect, for now we have *capacitive* stray currents between the two conductors. However, now the voltage across the load will also differ from that at the generator, in spite of the line conductors being perfect. This is due to small inductive voltage drops across short segments of the line; we know that a line segment of length $\Delta z$ has an inductance $L' \Delta z$ ($L'$ is the line inductance per unit length). Of course, a real transmission line also has a resistance per unit length (due to imperfect conductors), so in addition we will have a resistive voltage drop across segments of the line.

Shown in Table 18.1 are the parameters $C'$, $G'$, $L'$, and $R'$ of the three mentioned transmission-line types. Note that most frequently $\mu = \mu_0$, that the conductivity of the conductors is approximately $\sigma_c \simeq 56 \times 10^6$ S/m (copper), that the relative permittivity of the dielectric is usually 1.0 (air) or 2.1 to 4.0 (most other dielectrics, although dielectrics with considerably higher relative permittivity are also used), and that the conductivity of the dielectrics other than air is on the order of $10^{-12}$ S/m.

Thus if we wish to analyze any transmission line with ac excitation by circuit-theory concepts, we need to represent it as a series connection of many small cells containing series inductors and resistors, and parallel capacitors and resistors, as in Fig. 18.2. Such circuits are said to have *distributed parameters*. If losses in a line are very small, the line is referred to as a *lossless line*. Although all lines have losses, they can frequently be neglected, so analysis of lossless lines is of considerable practical interest.

TABLE 18.1 Parameters of Some Transmission Lines at High Frequencies

| Parameter | Coaxial line | Two-wire line ($d \gg 2a$) | Strip line ($b \gg a$) |
|---|---|---|---|
| | | $2a \qquad 2a$ $\epsilon, \mu, \sigma_d$ $d$ | $a$ $\epsilon, \mu, \sigma_d$ $b$ |
| $C' \left( \dfrac{\text{F}}{\text{m}} \right)$ | $\dfrac{2\pi\epsilon}{\ln b/a}$ | $\dfrac{\pi\epsilon}{\ln d/a}$ | $\epsilon \dfrac{b}{a}$ |
| $G' \left( \dfrac{\text{S}}{\text{m}} \right)$ | $\dfrac{\sigma_d}{\epsilon} C'$ | $\dfrac{\sigma_d}{\epsilon} C'$ | $\dfrac{\sigma_d}{\epsilon} C'$ |
| $L'_{ext}{}^* \left( \dfrac{\text{H}}{\text{m}} \right)$ | $\dfrac{\mu}{2\pi} \ln \dfrac{b}{a}$ | $\dfrac{\mu}{\pi} \ln \dfrac{d}{a}$ | $\mu \dfrac{a}{b}$ |
| $R' \left( \dfrac{\Omega}{\text{m}} \right)$ | $\dfrac{R_s}{2\pi} \left( \dfrac{1}{a} + \dfrac{1}{b} \right)$ | $\dfrac{R_s}{\pi a}$ | $\dfrac{2R_s}{b}$ |

$*L'_{int} = R'/\omega$ in all three cases, and $R_s = \sqrt{\omega\mu/2\sigma_c}$ (see Examples 21.7 and 21.9 for proof); $\sigma_c$ is the conductivity of the line conductors. PRooFs in Ex 20.4 - 20.6.
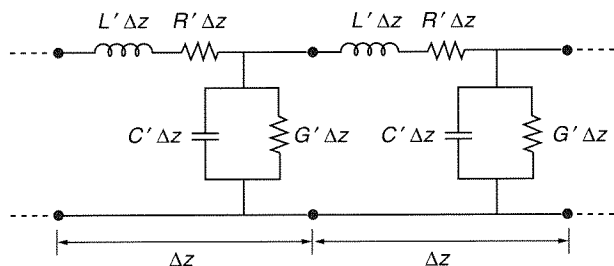


Figure 18.2 Circuit-theory approximation of a transmission line with losses

We will develop first the theory for the analysis of lossless lines and then introduce losses in a simple manner. The analysis will show that the time-varying voltage and current along the line vary continuously and that these variations propagate along the line. These are known as *voltage* and *current waves*. The analysis will also show that the voltages, currents, and the ratio of voltage and current at a point along the line depend on what load is connected at the end of the line. Typically, we wish to efficiently deliver power from a generator, through a line, to the load.

*Questions and problems:* P18.1 and P18.2

## 18.2 Analysis of Lossless Transmission Lines

The circuit-theory approximations of short and long sections of a lossless transmission line are sketched in Figs. 18.3a and b. Consider the three short sections in
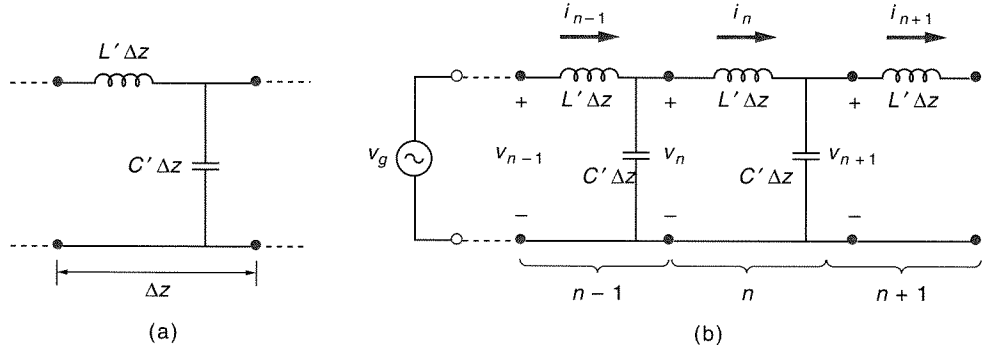
Figure 18.3 (a) A very short piece of a lossless transmission line of length $\Delta z$ can be represented as a circuit consisting of a series inductor and a shunt (parallel) capacitor. (b) A longer piece of the line can be represented as many cascaded short sections.

Fig. 18.3b, labeled $(n-1)$, $n$, and $(n+1)$. Let us apply Kirchhoff's voltage and current laws. The voltage across the $n$th inductor and the current through the $n$th capacitor are

$$\Delta L \frac{di_n}{dt} = v_n - v_{n+1} \quad \text{and} \quad \Delta C \frac{dv_n}{dt} = i_{n-1} - i_n. \tag{18.1}$$

Dividing both equations by $\Delta z$ and noting that

$$\frac{\Delta L}{\Delta z} = L' \quad \text{and} \quad \frac{\Delta C}{\Delta z} = C', \tag{18.2}$$

we can rewrite Eqs. (18.1) as follows:

$$L' \frac{di_n}{dt} = -\frac{v_{n+1} - v_n}{\Delta z} \quad \text{and} \quad C' \frac{dv_n}{dt} = -\frac{i_n - i_{n-1}}{\Delta z}. \tag{18.3}$$

As $\Delta z$ approaches zero, the right-hand sides become derivatives with respect to the coordinate $z$ (note that the left-hand sides are true derivatives with respect to time), and Eqs. (18.3) become

$$\frac{\partial v(t, z)}{\partial z} = -L' \frac{\partial i(t, z)}{\partial t} \quad \text{and} \quad \frac{\partial i(t, z)}{\partial z} = -C' \frac{\partial v(t, z)}{\partial t}. \tag{18.4}$$

*(Transmission-line equations, or telegraphers' equations, for lossless lines)*

Partial derivatives need to be used because $u(z, t)$ and $i(z, t)$ are functions of time, $t$, and distance along the line, $z$. It is clear that if voltage and current vary in *time*, they also vary *along the line*. Equations (18.4) are called the *transmission-line equations* or the *telegraphers' equations*. These two equations are coupled differential equations in two unknowns, $i$ and $v$.

It is easy to obtain instead equations with only voltage or only current. To that aim, take the derivative with respect to $z$ ($\partial/\partial z$) of the first equation and the time derivative ($\partial/\partial t$) of the second equation and eliminate the current (or voltage) by

substitution. Following this procedure, we obtain

$$\frac{\partial^2 v(t, z)}{\partial t^2} - \frac{1}{L'C'}\frac{\partial^2 v(t, z)}{\partial z^2} = 0 \qquad \frac{\partial^2 i(t, z)}{\partial t^2} - \frac{1}{L'C'}\frac{\partial^2 i(t, z)}{\partial z^2} = 0. \qquad (18.5)$$

*(Wave equations for voltage and current along lossless transmission lines)*

These equations are known as the *wave equations*. They describe the variation of voltage and current along a line and in time. The same or similar type of equation can be used to describe the electric and magnetic fields in a radio wave or optical ray, sound waves in acoustics, etc. We will later derive the same equation for the electric and magnetic field strength vectors, **E** and **H**, instead of voltages and currents.

### 18.2.1    FORWARD AND BACKWARD VOLTAGE WAVES IN THE TIME DOMAIN

Consider the voltage wave equation. It is not difficult to show (see Example 18.1) that its solution is

$$v(t, z) = V_+ f(t - z/c) + V_- g(t + z/c) \qquad \text{(V)}, \qquad (18.6)$$

*(Forward and backward voltage wave on a transmission line)*

where $V_+$ and $V_-$ are constants, $f$ and $g$ are *arbitrary* functions of the indicated arguments, and

$$c = \frac{1}{\sqrt{L'C'}} \qquad \text{(m/s)}. \qquad (18.7)$$

*(Velocity of a wave propagating along a transmission line)*

The physical meaning of the solution in Eq. (18.6) is as follows. Consider the function $f(t, z) = f(t - z/c)$ (the constant $V_+$ is irrelevant). Let the function at $t = 0$ be as $f(0, z)$ in Fig. 18.4. At a somewhat later instant, say $t = \Delta t$, the difference $t - z/c$ will have the same value as for $t = 0$ if we consider a point $z + c\,\Delta t = z + \Delta z$ instead of point $z$. This means that the bell-shaped voltage pulse $f(0, z)$ will have exactly the same form, but will be moved by $\Delta z = c\Delta t$, as indicated by the pulse labeled $f(\Delta t, z + c\,\Delta t)$ in Fig. 18.4. Because $\Delta t$ is arbitrary, this means that the voltage pulse moves from left to right in Fig. 18.4 (i.e., in the direction of the z axis) with a velocity $c = 1/\sqrt{L'C'}$. The wave moving in the $+z$ direction is the *forward traveling (voltage) wave* or the *incident (voltage) wave*.

It is simple to conclude in the same way that the function $g(t + z/c)$ represents a voltage wave propagating in the $-z$ direction. Such a wave is the *backward traveling (voltage) wave* or the *reflected (voltage) wave*.
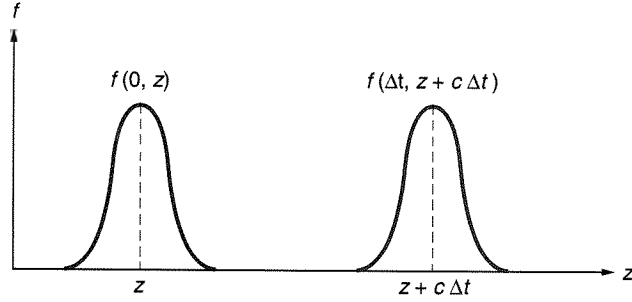
Figure 18.4 A voltage wave moves unchanged in shape,
with constant velocity $c$, along a lossless transmission line.

What is the velocity of propagation of voltage waves along a typical cable? For a 50-$\Omega$ coaxial cable with a typical dielectric, $C' \simeq 1\,\text{pF/cm}$, and $L' \simeq 2.5\,\text{nH/cm}$. The velocity $c$ in Eq. (18.7) for these $C'$ and $L'$ is about two-thirds of the speed of light in air (i.e., about $2 \times 10^8\,\text{m/s}$).

**Example 18.1—Proof that $f(t - z/c)$ and $g(t + z/c)$ are solutions of the wave equation.** The proof is simple if we recall the rules for finding the derivatives of a function of several variables. Suppose we have a function $f(x)$, where $x = x(t, z)$ is an arbitrary function of two independent variables, $t$ and $z$. The partial derivative of $f(x)$ with respect to $t$, for example, is obtained using the chain rule as follows:

$$\frac{\partial f(x)}{\partial t} = \frac{\partial f(x)}{\partial x} \frac{\partial x(t, z)}{\partial t}.$$

The second partial derivative is obtained in an analogous manner.
Let $x(t, z) = (t \pm z/c)$. Then

$$\frac{\partial f(x)}{\partial t} = \frac{\partial f(x)}{\partial x} \frac{\partial(t \pm z/c)}{\partial t} = \frac{\partial f(x)}{\partial x}, \tag{18.8a}$$

because $\partial(t \pm z/c)/\partial t = 1$. Hence also

$$\frac{\partial^2 f(x)}{\partial t^2} = \frac{\partial^2 f(x)}{\partial x^2}.$$

The derivative with respect to $z$ is somewhat different because $z$ is multiplied by a constant, $\pm 1/c$:

$$\frac{\partial f(x)}{\partial z} = \frac{\partial f(x)}{\partial x} \frac{\partial(t \pm z/c)}{\partial z} = \pm \frac{1}{c} \frac{\partial f(x)}{\partial x}, \tag{18.8b}$$

and

$$\frac{\partial^2 f(x)}{\partial z^2} = \frac{1}{c^2} \frac{\partial^2 f(x)}{\partial x^2}.$$

Substituting the second derivatives with respect to $t$ and $z$ into the wave equation (18.5), we see that it is indeed satisfied for *any* function $f(t \pm z/c)$.

Note that according to Eqs. (18.8a) and (18.8b),

$$\frac{\partial f(x)}{\partial z} = \pm \frac{1}{c} \frac{\partial f(x)}{\partial x} = \pm \frac{1}{c} \frac{\partial f(x)}{\partial t}. \tag{18.9}$$

## 18.2.2  FORWARD AND BACKWARD VOLTAGE WAVES IN THE COMPLEX (FREQUENCY) DOMAIN

Here we deal mostly with sinusoidally time-varying voltages and currents and linear materials, which means that we can use phasor (complex) notation. Both voltage and current have an assumed exponential time variation,

$$v, i \propto e^{j\omega t} \qquad j = \sqrt{-1} \quad \text{(the imaginary unit)}, \tag{18.10}$$

but these exponentials cancel out when we write equations involving $V$ and $I$ in phasor form. We use capital letters for rms (root mean square) complex quantities. The derivative with respect to time becomes just a multiplication with $j\omega$. Consequently, we can write the transmission-line equations (18.4) for sinusoidal time variation as

$$\frac{dV(z)}{dz} = -j\omega L' I(z) \qquad \text{and} \qquad \frac{dI(z)}{dz} = -j\omega C' V(z). \tag{18.11}$$

*[Lossless-transmission-line equations in phasor (complex) form]*

Eliminating the current from these equations results in

$$\frac{d^2 V(z)}{dz^2} = -\omega^2 L' C' V(z) = (j\omega \sqrt{L'C'})^2 V(z) = (j\beta)^2 V(z). \tag{18.12}$$

*[Voltage wave equation along lossless transmission lines in phasor (complex) form]*

The solution to this second-order differential equation is of the form

$$V(z) = V_+ e^{-j\beta z} + V_- e^{+j\beta z} \qquad \text{(V)}, \tag{18.13}$$

*(Total voltage wave in phasor notation)*

where

$$\beta = \omega \sqrt{L'C'} = \frac{\omega}{c} \qquad \text{(1/m)}. \tag{18.14}$$

*(Definition of phase constant)*

The constant $\beta$ is known as the *phase constant* (or *phase coefficient*) because it determines the phase of the voltage at a distance $z$ from the origin ($z = 0$). Comparing Eq. (18.13) with Eq. (18.6), it can be inferred that $V_+e^{-j\beta z}$ is the complex representation of a forward traveling wave and $V_-e^{+j\beta z}$ that of a backward traveling wave.

### 18.2.3  WAVELENGTH ALONG TRANSMISSION LINES

Consider the expression for the forward traveling cosine voltage wave in the time domain,

$$v_+(t, z) = V_+\sqrt{2}\,\cos(\omega t - \beta z).$$

The argument of the cosine function remains the same if any multiple of $2\pi$ is added to it, or by moving by $\beta\,\Delta z = n \cdot 2\pi$, $n = \pm 1, \pm 2, \ldots$ along the line. The smallest distance $\Delta z = \lambda$ for which this happens is obtained from the equation $\beta\lambda = 2\pi$, from which we derive

$$\lambda = \frac{2\pi}{\beta} = \frac{2\pi}{(\omega/c)} = \frac{c}{f} \qquad \text{(m)}. \tag{18.15}$$

*(Definition of wavelength of sinusoidal waves)*

This distance, $\lambda$, is known as the *wavelength* of the sinusoidal wave.

Note again that in complex (phasor) notation, the forward traveling wave has a minus sign in the exponential. This means that *for a fixed moment in time* the phase of the wave lags along the $z$ direction. (Because the wave is propagating in that direction, this must be the case.)

### 18.2.4  CURRENT WAVES IN THE COMPLEX (PHASOR) DOMAIN, AND THE CHARACTERISTIC IMPEDANCE

Expressions analogous to those for the voltage wave can be obtained for the current wave along the line. From Eqs. (18.4), (18.6), and (18.9), we find

$$\frac{\partial i(t, z)}{\partial t} = -\frac{1}{L'}\frac{\partial v(t, z)}{\partial z} = \frac{1}{cL'}\left[V_+\frac{\partial f(t - z/c)}{\partial t} - V_-\frac{\partial g(t + z/c)}{\partial t}\right].$$

This equation can be integrated directly with respect to time. Assuming zero dc components of voltages and currents and having in mind Eq. (18.7), the integration results in

$$i(t, z) = \frac{V_+}{\sqrt{L'/C'}}f(t - z/c) - \frac{V_-}{\sqrt{L'/C'}}g(t + z/c) \qquad \text{(A)}. \tag{18.16}$$

*(Forward and backward current waves along transmission lines)*

In phasor notation this equation becomes

$$I(z) = \frac{V_+}{Z_0} e^{-j\beta z} - \frac{V_-}{Z_0} e^{+j\beta z} \qquad \text{(A)}, \tag{18.17}$$

*(Forward and backward current waves in phasor notation)*

where

$$Z_0 = \sqrt{\frac{L'}{C'}} \qquad (\Omega). \tag{18.18}$$

*(Characteristic impedance of lossless line)*

$Z_0$ is called the *characteristic impedance* of the lossless transmission line. Like $L'$ and $C'$, it depends only on how the line is built (its dimensions and the materials used in it).

**Example 18.2—Numerical values of $c$ and $Z_0$ for some lossless transmission lines.** Because for lossless lines $L'_{\text{int}} = 0$, $R' = 0$, and $G' = 0$, for the three lines given in Table 18.1 the velocity of propagation, $c$, becomes

$$c = \frac{1}{\sqrt{L'C'}} = \frac{1}{\sqrt{\epsilon\mu}} \qquad \text{(for the three lines in Table 18.1).}$$

It can be shown that this simple relation is valid not only for the three lines considered but for all lossless transmission lines with a homogeneous dielectric.

In particular, if the dielectric in the line is air we have

$$c_0 = \frac{1}{\sqrt{\epsilon_0\mu_0}} = \frac{1}{\sqrt{8.854 \cdot 10^{-12} \times 4\pi \cdot 10^{-7}}} \simeq 3 \times 10^8 \ \text{m/s},$$

i.e., the velocity of propagation of voltage and current waves along air lines equals the velocity of light in a vacuum. This is a conclusion to be remembered. Note that it is valid only for lines with air dielectric. Because for any dielectric $\epsilon > \epsilon_0$, the propagation velocity along lines with dielectric other than air is always less than the velocity of light in a vacuum.

The characteristic impedance, $Z_0$, is different for the three lines. Let us write the explicit expression for the coaxial line:

$$Z_0 = \sqrt{\frac{L'}{C'}} = \frac{1}{2\pi} \sqrt{\frac{\mu}{\epsilon}} \ln \frac{b}{a} \quad \text{(lossless coaxial line).}$$

For example, if $b/a = e = 2.71828$ and the dielectric is air, we obtain that $Z_0 \simeq 60 \ \Omega$. Characteristic impedance of commercial coaxial lines (for which $\epsilon > \epsilon_0$) ranges from about $50 \ \Omega$ to about $90 \ \Omega$.

Generally, we have both a forward and a backward wave on the line. To calculate the ratio $v(t, z)/i(t, z)$ at any point along the line and at any instant, the complete expressions for $v(t, z)$ and $i(t, z)$ in Eqs. (18.6) and (18.16) must be used. If only a forward wave exists along a line, the ratio of the forward voltage and current waves is found from Eqs. (18.6) and (18.16) to be

$$\frac{v(t, z)}{i(t, z)} = Z_0. \tag{18.19}$$

*(Only a forward wave along the line)*

This means that if only a forward wave exists along the line, the ratio of voltage and current at *any* point along the line and at *any* instant of time is the same, equal to $Z_0$.

If only the backward wave propagates along the line, Eqs. (18.6) and (18.16) yield

$$\frac{v(t, z)}{i(t, z)} = -Z_0. \tag{18.20}$$

*(Only a backward wave along the line)*

These two equations have a simple physical meaning. Consider Fig. 18.5a and assume reference directions of voltage $v_+$ and current $i_+$ as indicated. If only a forward wave exists, the generator is at the left in Fig. 18.5a, and the line to the right is such that no backward (reflected) wave is created. This will be the case if the line is infinitely long to the right. This means that a generator connected to the input terminals of an infinitely long lossless line will see the line as a resistor of resistance equal to $Z_0$.

The last conclusion enables us to understand an extremely important fact: because an infinite section of any lossless line with respect to its input terminals behaves as a resistor of resistance $Z_0$, we can eliminate the reflected wave on a line of *any* length by terminating the line in its characteristic impedance. If this is done, we say that the line is *matched*.

Why do we have the minus sign in Eq. (18.20)? Note that the reference directions of voltage and current have been adopted as indicated in Figs. 18.5a and b (the reference conductor for voltage is designated by a "+" sign and the reference direc-
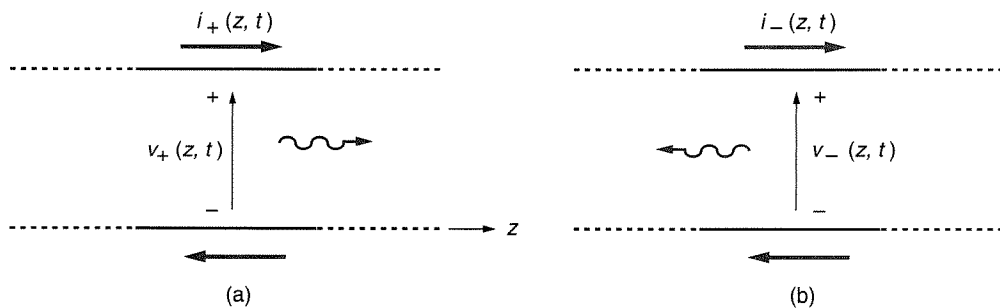


Figure 18.5 (a) Forward and (b) backward voltage and current waves in a transmission line. Note that the adopted reference directions of voltage and current are the same in both cases.

tion for current by an arrow). If the backward wave alone propagates along the line, the generator must be at the right end of the line, feeding an infinite line extending to the left. In Fig. 18.5b the same reference directions are adopted for voltage and current as in Fig. 18.5a; therefore one of these quantities for a reflected wave must be negative so that the power flow is from right to left. Because we retained the same sign for the backward voltage wave, the current wave must change sign with respect to the forward wave. The meaning of the negative sign is exactly the same as in circuit theory: the current is in the direction opposite to the reference direction.

Assume that there is a backward (reflected) wave along the line. Let the generator be connected at the line end toward which the backward wave is propagating (to the left in Fig. 18.5b). When the backward wave reaches the generator, will it produce a backward-backward (i.e., a new forward) wave? The answer is evident: such a wave *will* be produced unless the internal resistance of the generator equals the line characteristic impedance, $Z_0$. For this reason generators (or equivalent Thévenin's generators), if possible, are made to have this internal resistance, usually 50 $\Omega$. In what follows, we assume that generators driving transmission lines satisfy this condition, i.e., that they are matched to the line.

*Questions and problems:*    Q18.1 to Q18.11, P18.3 to P18.7

## 18.3  Analysis of Terminated Lossless Transmission Lines in Frequency Domain

The excitation of transmission lines can have any time variation. Frequently the excitation is sinusoidal or nearly sinusoidal. In this and the next section we restrict our attention to sinusoidal voltages and currents along transmission lines and use the phasor (complex) notation. This section is devoted to lossless lines, and the next to lines with losses.

In reality, a line may be terminated in *any* load, which is not necessarily the same as its characteristic impedance, as shown in Fig. 18.6. We now know that a forward (or incident) wave travels from the generator to the right in Fig. 18.6. When
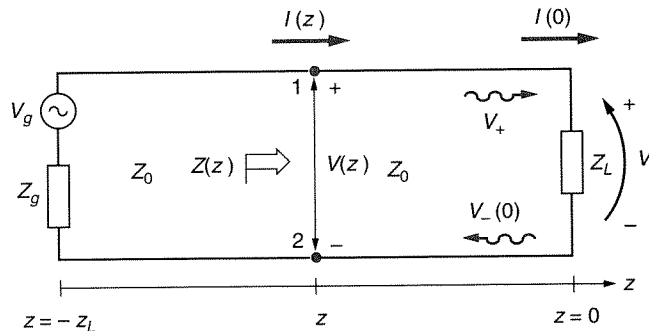


Figure 18.6 A transmission line of characteristic impedance $Z_0$ terminated in a load $Z_L$ at a distance $z_L$ from the generator of internal impedance $Z_0$. The coordinate origin, $z = 0$, is adopted at the position of the load.

it reaches the load, some of the power is absorbed, and some is reflected, giving rise to a backward wave. The line being linear (the fundamental transmission-line equations are linear), the amplitude of the reflected wave is proportional to that of the incident wave. Because we normally assume that the generator is matched to the line, when the reflected wave reaches the generator it is absorbed in its internal impedance.

The coordinate origin, $z = 0$, can be anywhere along the line. In the analysis of transmission lines, we are mostly concerned with the load because this is where we wish the generator power to be delivered. Therefore, it is convenient to shift the origin from the generator to the load, as in Fig. 18.6.

### 18.3.1 THE REFLECTION AND TRANSMISSION COEFFICIENTS

The *(voltage) reflection coefficient* is defined as the ratio of the complex amplitudes (or rms values) of the reflected and incident voltage waves at the load. If $z = 0$ is at the load, as in Fig. 18.6, the reflection coefficient is given by

$$\rho = \frac{V_-}{V_+} \quad \text{(dimensionless).} \tag{18.21}$$

*(Definition of the reflection coefficient)*

With this definition, the phasor voltage and current along the line in Eqs. (18.13) and (18.17) can be written in the form

$$V(z) = V_+ e^{-j\beta z}(1 + \rho e^{2j\beta z}) \quad \text{(V),} \tag{18.22a}$$

$$I(z) = \frac{V_+}{Z_0} e^{-j\beta z}(1 - \rho e^{2j\beta z}) \quad \text{(A).} \tag{18.22b}$$

*(Total voltage and current along a transmission line in terms of the reflection coefficient)*

When a generator is connected at one end of a line and a load at the other end, part of the power is reflected from the load (if the load is not perfectly matched) and part of the power is delivered to the load. Generally the goal is to deliver as much power to the load as possible. A quantity that describes the voltage across the load as a function of the incident voltage is called the *transmission coefficient*, defined by

$$\tau = \frac{V_{\text{load}}}{V_+} = \frac{V_+ + V_-}{V_+} = 1 + \rho \quad \text{(dimensionless).} \tag{18.23}$$

*(Definition of transmission coefficient)*

The magnitude of the voltage reflection coefficient for passive loads is smaller than or equal to unity, whereas the magnitude of the transmission coefficient is smaller than or equal to 2. The following examples illustrate this range of values.

**Example 18.3—Reflection and transmission coefficients for shorted, open, and matched transmission lines.** Let us look at a few simple and extreme examples of terminations (loads) in Fig. 18.6: (1) an open circuit ($Z_L = \infty$), (2) a short circuit ($Z_L = 0$), and (3) a matched load ($Z_L = Z_0$).

At an open end of a line, no current flows between the two conductors. As the adopted reference directions of forward and backward current waves are the same (Figs. 18.5a and b) the reflected current at that point has to be of the same magnitude as the incident current wave, but of opposite sign. According to Eqs. (18.13) and (18.17), the reflected voltage wave at that point is then equal to the incident wave (note reference directions for the two voltages in Fig. 18.5). Consequently, the voltage reflection coefficient at the load for an open end is $\rho = 1$. From Eq. (18.22a) it follows that at the open-circuited line end the total voltage is twice the incident voltage, corresponding to a voltage transmission coefficient $\tau = 2$.

At a short-circuited line end, there is no voltage between the two line conductors, corresponding to a transmission coefficient $\tau = 0$. Referring to reference directions for voltage in Fig. 18.5, as the total voltage at the end of the line has to be zero, the reflected voltage is the negative of the incident voltage. The (voltage) reflection coefficient for a zero load is therefore $\rho = -1$. According to Eq. (18.22b), the current at the short-circuited end is twice the current of the incident current wave.

For a matched case (load impedance equal to the line characteristic impedance), if we divide Eq. (18.22a) by Eq. (18.22b) and set $z = 0$ this ratio is equal to the load impedance, in this case $Z_0$. So we obtain the following equation:

$$Z_0 = Z_0 \frac{1+\rho}{1-\rho}.$$

This equation can be satisfied only if $\rho = 0$. This was to be expected because we know that a matched load absorbs the incident wave completely, corresponding to a transmission coefficient of $\tau = 1$.

**Example 18.4—Time-average power delivered to the load.** From circuit theory we know that the time-average power delivered to a load is obtained from the phasor voltage across the load, $V$, and the phasor current in the load, $I$, as $P_{av} = \mathrm{Re}\{V \cdot I^*\}$, where the asterisk denotes a complex conjugate. The voltage and current across the load are obtained from Eqs. (18.22a) and (18.22b) if we set $z = 0$. Thus the average power delivered to the load terminating a transmission line is

$$P_{\text{load av}} = \mathrm{Re}\{V(0)I^*(0)\} = \mathrm{Re}\left\{ \frac{|V_+|^2}{Z_0}[1 - |\rho|^2 + (\rho - \rho^*)] \right\}.$$

Recall that for a complex number $a = b + jc$, $a - a^* = (b + jc) - (b - jc) = j2c$. Therefore $(\rho - \rho^*)$ is purely imaginary, so that

$$P_{\text{load av}} = \frac{|V_+|^2}{Z_0}(1 - |\rho|^2). \tag{18.24}$$

*(Average power delivered to a transmission-line load)*

Note that this is precisely the difference of average power of the incident wave, $|V_+|^2/Z_0$, and of the reflected wave, $|V_-|^2/Z_0 = |\rho|^2|V_+|^2/Z_0$.

Usually we wish to express the power delivered to the load in terms of the voltage $V_g$ of the generator connected at the input transmission-line terminals. As

mentioned, we always assume that the generator is matched to the line, i.e., that its internal impedance is equal to the line characteristic impedance. Because for the incident voltage wave we assume an infinite line, the impedance of the line seen by the *incident wave* at the generator terminals is simply $Z_0$. Therefore, $V_+ = V_g/2$ for a matched generator.

### 18.3.2   IMPEDANCE OF A TERMINATED TRANSMISSION LINE

Consider again the cross section of the line at $z$ in Fig. 18.6. Looking to the right from points 1 and 2, we have a passive network (containing no generators) with two terminals (points 1 and 2). Considering it as a black box, we define its impedance in the usual manner as the ratio of the voltage between the terminals (which is the *total* voltage) and the corresponding current (which is the *total* current). (Note that the adopted reference directions of voltage and current in Fig. 18.6 are precisely as needed for an impedance element to the right of points 1 and 2.) This impedance is a function of $z$. According to Eqs. (18.22$a$) and (18.22$b$), we have

$$Z(z) = \frac{V(z)}{I(z)} = Z_0 \frac{1 + \rho e^{j2\beta z}}{1 - \rho e^{j2\beta z}}.$$

This is the impedance of the line looking toward the load at a distance $z$ from the coordinate origin (assumed at the load). Due to the adopted coordinate origin, the $z$ coordinate of any point is negative (Fig. 18.7). To avoid the minus sign in the expressions to follow, it is convenient to introduce a new coordinate, $\zeta = -z$ (Fig. 18.7), representing the distance from the load. With this change in variable along the line, the expression for the impedance in the last equation becomes

$$Z(\zeta) = Z_0 \frac{1 + \rho e^{-j2\beta\zeta}}{1 - \rho e^{-j2\beta\zeta}}. \tag{18.25}$$

In particular, for $\zeta = 0$ we have that $Z(0) = Z_L$. Thus

$$Z_L = Z_0 \frac{1 + \rho}{1 - \rho}, \tag{18.26}$$
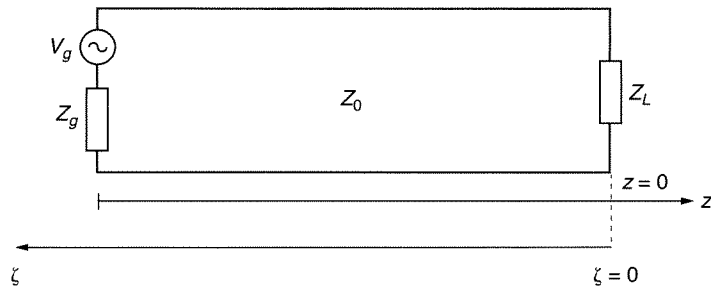


Figure 18.7 Coordinate origin, $z = 0$, at the load, and the coordinate $\zeta = -z$

which is used for determining $Z_L$ if $\rho$ has been determined experimentally. Solving the last equation, we can obtain expressions for $\rho$ and $\tau$ as a function of the load impedance:

$$\rho = \frac{Z_L - Z_0}{Z_L + Z_0} \quad \text{(dimensionless),} \tag{18.27a}$$

*(Alternative expression for reflection coefficient)*

$$\tau = \frac{2Z_L}{Z_L + Z_0} \quad \text{(dimensionless),} \tag{18.27b}$$

*(Alternative expression for transmission coefficient)*

so that the reflection and transmission coefficients can be determined knowing only the load impedance and the line characteristic impedance.

Finally, if we substitute $\rho$ from Eq. (18.27a) into Eq. (18.25) and recall that $e^{j\alpha} = \cos\alpha + j\sin\alpha$ (Euler's formula), the input impedance of a section of line of length $\zeta$ terminated in $Z_L$, Eq. (18.25), after simple manipulations becomes

$$Z(\zeta) = Z_0 \frac{Z_L \cos\beta\zeta + jZ_0 \sin\beta\zeta}{Z_0 \cos\beta\zeta + jZ_L \sin\beta\zeta} = Z_0 \frac{Z_L + jZ_0 \tan\beta\zeta}{Z_0 + jZ_L \tan\beta\zeta} \quad (\Omega). \tag{18.28}$$

*(Input impedance of line of length $\zeta$ terminated in impedance $Z_L$)*

The important thing to remember is that the characteristic impedance $Z_0$ depends only on the way the line is made. The impedance along the line (looking toward the terminating impedance) is quite different: it depends on both $Z_0$ and the terminating impedance but also on the coordinate along the line.

**Example 18.5—Input impedance of an open line.** Assume that the line is open. This corresponds to $Z_L = \infty$ in Eq. (18.28), so that the input impedance of a section of the line of length $\zeta$ becomes

$$Z(\zeta) = -j\frac{Z_0}{\tan\beta\zeta} = -jZ_0 \overset{cotan}{\underline{\phantom{tan}}} \beta\zeta.$$

If $\beta\zeta < \pi/2$, that is, if $\zeta < \pi/2 \cdot \lambda/(2\pi) = \lambda/4$, $Z(\zeta)$ is a negative imaginary number. This means that the line behaves as a capacitor. (Note, however, that this line behavior is valid only for a line length less than $\lambda/4$!)

You might recall from Chapter 2 that the parasitic inductance of rf chip capacitors makes these elements look predominantly inductive above a certain frequency (the lead inductance is on the order of 1 nH). At microwave frequencies, short sections of open-ended lines are frequently used to obtain in a simple manner a capacitive reactance of a desired value at a given frequency. Note that this reactance depends on frequency in a different way than in the case of a capacitor (for which $X_C = -1/\omega C$).

As an example, consider a short segment of length 1 cm of a 50-$\Omega$ line. Assume that the velocity of wave propagation along the line is $0.67c_0$. The reactance of this line segment at $f = 1000$ MHz is

$$Z(1 \text{ cm})_{1000\text{MHz}} = -jZ_0 \overset{\text{COTAN}}{\cancel{\tan}}{}^{-1}\left(\frac{2\pi}{c/f} \times \zeta\right)$$

$$= -j50 \overset{\text{COTAN}}{\cancel{\tan}}{}^{-1}\left(\frac{2\pi}{0.67 \cdot 3 \cdot 10^8/10^9} \times 0.01\right) \simeq -j320\Omega,$$

which corresponds to a capacitance of $1/(2\pi \cdot 10^9 \cdot 320) \simeq 0.5$ pF (only at 1000 MHz!). It is suggested as an exercise for the reader to calculate the capacitance of the line between 900 MHz and 1100 MHz.

If $\lambda/4 < \zeta < \lambda/2$, the line behaves as an inductive element; for a still greater length it behaves again as a capacitive element, and so on. It is left as an exercise for the reader to plot $Z(\zeta)$ as (1) a function of frequency and (2) function of the length of the line in wavelengths.

**Example 18.6—Input impedance of a shorted line.** Assume now that the line is shorted, i.e., that in Eq. (18.28) $Z_L = 0$. The input impedance of a section of the line of length $\zeta$ in this case is

$$Z(\zeta) = +jZ_0 \tan \beta\zeta.$$

If $\beta\zeta < \pi/2$, that is, if $\zeta < \pi/2 \cdot \lambda/(2\pi) = \lambda/4$, $Z(\zeta)$ is a *positive* imaginary number, i.e., the line behaves as an inductor.

You might also recall from Chapter 2 that an inductor has parasitic capacitance between the windings, and as the frequency increases the element looks more and more like a capacitor. Therefore it is hard to make inductors at microwave frequencies. Shorted sections of transmission lines are used frequently at microwave frequencies to obtain in a simple manner an inductive reactance of desired value. Note, however, that this reactance depends on frequency in a different way from that of an inductor ($X_L = \omega L$).

If $\lambda/4 < \zeta < \lambda/2$, the line behaves as a capacitive element; for a still greater length it behaves again as an inductive element, and so on. It is left as an exercise for the reader to plot $Z(\zeta)$ as (1) a function of frequency and (2) as a function of the length of the line in wavelengths.

**Example 18.7—Quarter-wave transformers.** An interesting and important case is when the length of the transmission line is a quarter of a wavelength. Because then $\beta\zeta = (2\pi/\lambda) \times (\lambda/4) = \pi/2$, from Eq. (18.28) we obtain

$$Z\left(\frac{\lambda}{4}\right) = \frac{Z_0^2}{Z_L}. \tag{18.29}$$

The load impedance is transformed from a value $Z_L$ to a value $Z_0^2/Z_L$. For example, if $Z_L$ is a high impedance, $Z$ will be low and vice versa. Quarter-wavelength transmission-line sections often play the same role at rf and microwave frequencies as impedance transformers at lower frequencies. (At high frequencies, it is difficult to build good transformers due to parasitic capacitances in inductors and also losses in the conductors and cores.)

Quarter-wave transformers are especially used for matching resistive loads. For example, if we want to match a 100-$\Omega$ load to a 50-$\Omega$ transmission line, we could use a quarter-wavelength section of a line with a characteristic impedance of $Z_0 = \sqrt{100 \cdot 50} = 70.7\ \Omega$.

However, unlike in a low-frequency transformer there is a phase lag in the section of the transmission line. This type of impedance transformer does not work for voltage and current transformation. Note also that the quarter-wavelength transformer effect works only in a narrow range of frequencies (it exactly works only at the frequency for which the length of the line is a quarter of a wavelength).

Analogous ideas are used in optics to make antireflection coatings for lenses. We explain this in more detail in later chapters.

**Example 18.8—Thévenin equivalent of an open-ended section of transmission line fed by a generator. Line terminated in an infinite line of different characteristic impedance.** Assume that a line of characteristic impedance $Z_1$ (line 1) is terminated in an infinite (or matched) line of characteristic impedance $Z_2$ (line 2). We know that line 2 from its input terminals represents a load of resistance $Z_2$. So line 1 can be regarded as terminated in a load $Z_L = Z_2$. Consequently we know the voltage and current distribution along line 1.

Along line 2 we have only a forward voltage and a forward current wave, propagating along it with a velocity determined by its capacitance and inductance per unit length. For determining these waves we need only determine the voltage at the input terminals of line 2. This can be done very simply by applying Thévenin's theorem.

Line 1 as seen from the input terminals of line 2 represents an equivalent real voltage generator. The voltage of the generator equals the open-circuit voltage at the end of line 1. From Example 18.3 we know that this voltage is twice the incident voltage along line 1, $V_{\text{Th}} = V_+(1 + \rho) = 2V_+$. The internal impedance of the Thévenin generator is the impedance of the infinite (or matched) line 1, so $Z_{\text{Th}} = Z_1$.

Let us summarize this very useful result. The equivalent Thévenin voltage source and impedance for an open-ended section of line of characteristic impedance $Z_1$ fed by a generator that gives an incident voltage $V_+$ are

$$Z_{\text{Th}} = Z_1 \qquad \text{and} \qquad V_{\text{Th}} = 2V_+.$$

After we replace line 1 with its Thévenin equivalent, the input voltage of line 2 can be found as in a voltage divider:

$$V_{2\ \text{input}} = V_{\text{Th}} \frac{Z_2}{Z_1 + Z_2} = 2V_+ \frac{Z_2}{Z_1 + Z_2}.$$

We know the reflection coefficient for line 1, given in Eq. (18.27a), which in this case becomes

$$\rho = \frac{Z_2 - Z_1}{Z_2 + Z_1}.$$

The transmission coefficient, from line 1 to line 2, is given by Eq. (18.27b):

$$\tau = \frac{V_{2\ \text{input}}}{V_+} = \frac{2Z_2}{Z_1 + Z_2}.$$

### 18.3.3   THE VOLTAGE STANDING-WAVE RATIO (VSWR)

A useful and frequently used concept related to the reflection coefficient is the *voltage standing-wave ratio*, or VSWR. The VSWR is the ratio of the maximal to minimal voltage along the line. Because $|e^{2j\beta z}| = 1$, according to Eq. (18.22a) the VSWR is given by

$$\text{VSWR} = \frac{V(z)_{\max}}{V(z)_{\min}} = \frac{1 + |\rho|}{1 - |\rho|}. \tag{18.30}$$

*(Definition of voltage standing-wave ratio, VSWR)*

Note that for a matched load, VSWR $= 1$, and for open and for short circuits, VSWR $= \infty$.

**Example 18.9—Standing waves on transmission lines.** When a line is matched at its end, we know that there is only a forward wave propagating along the line. To visualize such a voltage wave for sinusoidal excitation, imagine a sine function that moves along the line with a velocity $c$.

When the line is not matched there is another sinusoid, usually of smaller amplitude, moving from the load toward the generator (where we assume a matched load, i.e., no more reflected waves) with the same velocity, $c$. So in the general case we have two sine waves of unequal amplitudes moving in opposite directions with the same velocity. The total voltage at any point along the line (and at any moment) is obtained as their sum. Due to their equal velocities, however, there will be fixed minima and maxima of the total wave, as the following example shows.

We know that for a shorted line the voltage reflection coefficient $\rho = -1$ (see Example 18.3). Consequently, according to Eq. (18.22a), the total voltage along the line is of the form

$$V(z) = V_+ e^{-j\beta z}(1 - e^{2j\beta z}) = V_+(e^{-j\beta z} - e^{j\beta z}) = -j2V_+ \sin(\beta z),$$

because $e^{-j\alpha} - e^{j\alpha} = (\cos\alpha - j\sin\alpha) - (\cos\alpha + j\sin\alpha) = -j2\sin\alpha$.

The instantaneous value of the voltage along the line, $v(t, z)$, is hence obtained as

$$v(t, z) = \text{Re}\{-j2V_+\sqrt{2}\sin(\beta z)e^{j\omega t}\} = 2V_+\sqrt{2}\sin(\beta z)\sin\omega t.$$

This voltage does *not* have any argument of the form $(t \mp z/c)$! Consequently this is not a forward or a backward traveling voltage wave. Instead, it has zero values at all points where the sine has a zero value, and it oscillates *between* these fixed, stationary zeros of the total voltage. For this reason, this kind of wave is termed the *standing wave*. A sketch of the voltage standing wave for a sequence of time instants is shown in Fig. 18.8a. Note that a standing wave in phasor (complex) notation is easily recognized by the absence of the "traveling-wave" factor $e^{\mp j\beta z}$ (or any other coordinate instead of $z$).

The distance along two fixed, zero-voltage points along the line is given by

$$\Delta z_{\text{between two voltage zeros}} = \frac{\pi}{\beta} = \pi\frac{\lambda}{2\pi} = \frac{\lambda}{2}.$$
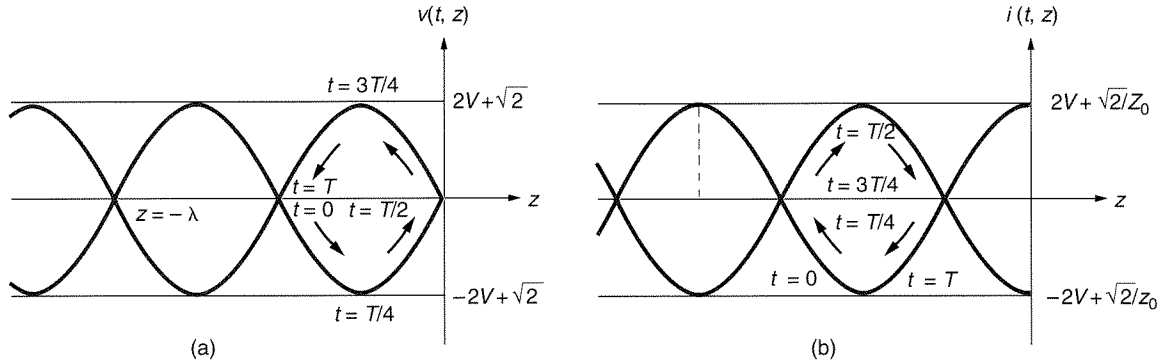
**Figure 18.8** (a) Voltage and (b) current standing wave along a shorted transmission line at indicated time instants (corresponding to the expressions derived in Example 18.9)

The total current along a shorted line has the same property, i.e., it is also a standing wave. From Eq. (18.22*b*), it is easily found that

$$i(t, z) = -2\frac{V_+}{Z_0}\sqrt{2}\cos(\beta z)\cos\omega t.$$

A sketch of the current standing wave for a sequence of time instants is shown in Fig. 18.8b.

It is left as an exercise for the reader to derive the expressions for standing voltage and current waves for an open transmission line.

Thus if we are able to measure the voltage along a transmission line we can easily conclude whether the line is shorted or open. The next example will show how we can measure the impedance of any load terminating a line by measuring, essentially, the VSWR.

**Example 18.10—Measurement of load impedance using a slotted line.** Figure 18.9a shows what is called the slotted coaxial line. Slotted lines may be used for measuring impedances at very high frequencies. To understand how this can be done, let the generator have a fixed but unknown frequency, and let the dielectric in the slotted line of characteristic impedance $Z_0$ be air. The coax is rigid and the outer conductor tube has a narrow slot along its length. The slot is made along the current-flow lines in the outer line conductor, so it only slightly affects the distribution of current and voltage in the line. A movable fixture is attached to the tube and contains a pinlike probe that protrudes through the slot and samples the electric field (voltage) inside the cable. Recall that the electric field vector in the cable is radial, so a voltage $v = \int_{\text{pin}} \mathbf{E} \cdot d\mathbf{l}$ is induced in the probe. This voltage is converted to a dc voltage by means of a diode detector and gives a measure of the relative electric field along the line. The position of the probe is measured along an arbitrary scale, e.g., like the one in Fig. 18.9b.

Usually the probe cannot slide all the way to the end on the line where the load is attached, and also the connector at the end of the line adds an unknown line length. So the first step in measuring an unknown load is to determine exactly where the line ends. We know that everything along a line is repeated every half wavelength. This means that we can determine the position of the end of the line displaced by an integer number of half wavelengths, so that it falls along our scale.
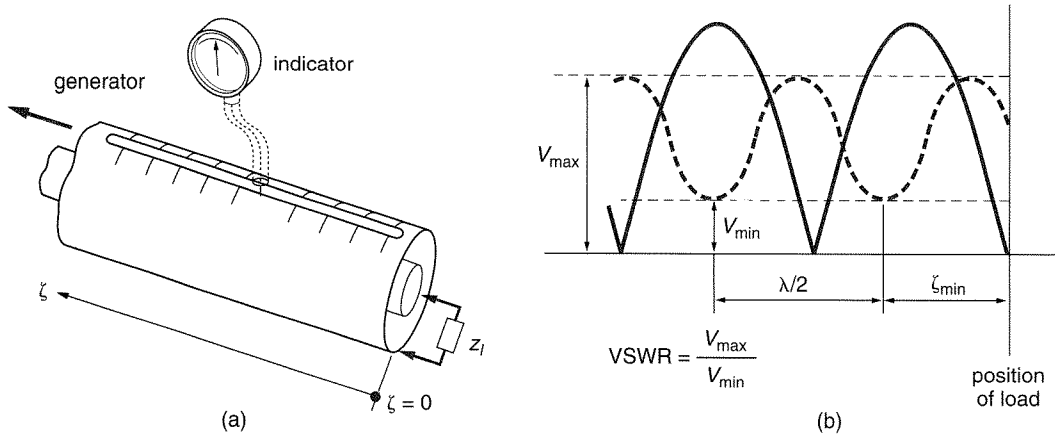
**Figure 18.9** (a) Sketch of a slotted coaxial line and (b) sketch of measured voltage along the line for a short circuit (solid line) and arbitrary complex load (dashed line)

How is the position of the end of the line determined? The easiest way is to connect a known load to the end, so that a standing wave is set up. If a short (or open) is connected, we know that the minima (maxima) of the standing wave occur at the load and every half wavelength away toward the generator. The wavelength measured along the line is practically equal to the free-space wavelength, $\lambda_0 = c_0/f$. Therefore we can measure the frequency of the generator simply by moving the probe back and forth and determining the two successive minima. Usually a short is used because minima are sharper and therefore more precise than maxima. (Sketch the derivative, or slope, of a standing wave to convince yourself.) The standing wave pattern due to a short is sketched in Fig. 18.9b in solid line, and the real and displaced positions of the end of the line are indicated.

After this calibration is performed, the unknown load is connected to the end of the line and the standing wave sampled once more. Again, two successive minima will be at a distance $\lambda_0/2$ apart, but they are displaced in position from the minima obtained with a short. This is because the phase of the load is different from that of the short. By moving the probe back and forth, we determine the maximum and minimum readings of the indicator, which gives us the voltage standing-wave ratio, VSWR, defined in Eq. (18.30), as sketched in Fig. 18.9b in dashed line. Then we locate as precisely as possible the distance $\zeta_{min}$ of the first minimum from the minimum obtained with a short, in terms of wavelength. From Eq. (18.22a) we find that the voltage minimum occurs when $\rho e^{-j2\beta\zeta}$ is real and negative, that is, equal to $-|\rho|$. So the impedance $Z(\zeta)$ given by Eq. (18.28) for $\zeta = \zeta_{min}$ is real, and equal to

$$Z(\zeta_{min}) = Z_0 \frac{1 - |\rho|}{1 + |\rho|} = \frac{Z_0}{\text{VSWR}}.$$

As we now know $Z(\zeta_{min})$ and $\zeta_{min}$ (in terms of wavelength), we can evaluate the unknown load impedance $Z_L$ from Eq. (18.28).

*Questions and problems:*    Q18.12 to Q18.16, P18.8 to P18.27

## 18.4   Lossy Transmission Lines

We know that a real transmission line has losses in the conductors (due to the finite conductivity of the metal) as well as in the dielectric between the conductors (due principally to the polarization losses in the dielectric). If the line is represented as a series connection of many short cells, these losses can be accounted for by a series and a shunt resistor in every cell, as in Fig. 18.10. The total series impedance per unit length is thus $R' + j\omega L'$ (instead of $j\omega L'$ for lossless lines), and the total shunt admittance per unit length is $G' + j\omega C'$ (instead of $j\omega C'$). The phasor equations (18.11) therefore take the form

$$\frac{dV(z)}{dz} = -(R' + j\omega L')I(z), \qquad \text{and} \qquad \frac{dI(z)}{dz} = -(G' + j\omega C')V(z). \qquad (18.31)$$

Noting that the lossless-line characteristic impedance in phasor form, $\sqrt{L'/C'}$, originally was $\sqrt{j\omega L'/j\omega C'}$, the characteristic impedance of a lossy line is given by

$$Z_0 = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}. \qquad (18.32)$$

*(Characteristic impedance of a lossy line)*

Similarly, the expression $j\beta = j\omega\sqrt{L'C'} = \sqrt{(j\omega L')(j\omega C')}$ in the exponential terms in Eq. (18.13) now becomes

$$\gamma = \alpha + j\beta = \sqrt{(R' + j\omega L')(G' + j\omega C')}. \qquad (18.33)$$

*(Propagation constant of a lossy line)*

The constant $\gamma$ is known as the *propagation constant* (or *propagation coefficient*) of the line, $\alpha$ as the *attenuation constant (coefficient)*, and $\beta$, as earlier, the *phase constant (coefficient)*.

Thus, for lossy lines and a forward wave, instead of $e^{-j\beta z}$ in the expressions for voltages and currents we now have $e^{-(\alpha + j\beta)z} = e^{-\alpha z}e^{-j\beta z}$. The factor $e^{-\alpha z}$ means that in addition to traveling in the $z$ direction, the amplitudes of the forward voltage and current waves also fall off in the direction of propagation. This is called *attenuation* and is a characteristic of every real transmission line. The phase of the wave is determined by $\beta$ (phase constant), and its attenuation by $\alpha$.
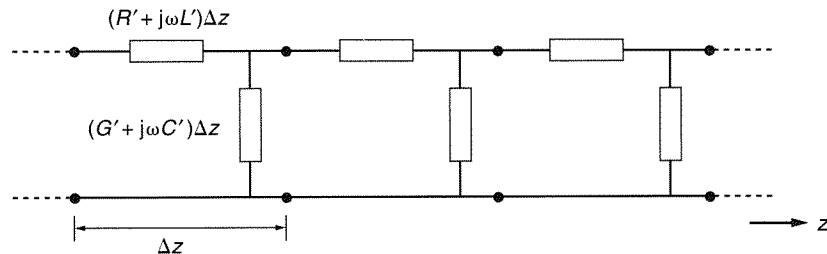


Figure 18.10 Schematic of a transmission line with distributed losses included.

Rearranging Eq. (18.33) we obtain

$$\gamma = \sqrt{j\omega L' j\omega C' \left(1 + \frac{R'}{j\omega L'}\right) \left(1 + \frac{G'}{j\omega C'}\right)} =$$

$$= j\omega\sqrt{L'C'}\sqrt{1 - j\left(\frac{R'}{\omega L'} + \frac{G'}{\omega C'}\right) - \frac{R'G'}{\omega^2 L'C'}}. \tag{18.34}$$

For transmission lines with small losses ($R' \ll \omega L'$ and $G' \ll \omega C'$), this can be written in approximate form

$$\gamma \simeq j\omega\sqrt{L'C'}\sqrt{1 - j\left(\frac{R'}{\omega L'} + \frac{G'}{\omega C'}\right)} \simeq j\omega\sqrt{L'C'}\left[1 - \frac{j}{2}\left(\frac{R'}{\omega L'} + \frac{G'}{\omega C'}\right)\right]. \tag{18.35}$$

So we find that for transmission lines with small losses

$$\alpha \simeq \frac{1}{2}\left(R'\sqrt{\frac{C'}{L'}} + G'\sqrt{\frac{L'}{C'}}\right) \qquad \beta \simeq \omega\sqrt{L'C'}. \tag{18.36}$$

*(Attenuation and phase constant for lines with small losses)*

If along a transmission line only the forward wave is propagating, both current and voltage along the line have the attenuation factor $e^{-\alpha z}$. Therefore the average power transmitted at a point $z$ in the direction of the wave, being the product of phasor rms voltage and conjugate current, is of the form $P(z) = P(0)e^{-2\alpha z}$.

If a quantity (e.g., voltage) at $z$ is of amplitude $V_+e^{-\alpha z}$, at $z + d$ it is of amplitude $V_+e^{-\alpha(z+d)}$. The attenuation of the voltage along this line section is frequently expressed as the natural logarithm of the ratio of the voltage amplitude at $z$ and that at $z + d$. The unit of this measure of attenuation is termed the *neper* (Np) [after the Scottish mathematician John Neper (Napier), who at the turn of the 16th century invented the logarithm]:

Attenuation of forward voltage wave in nepers

$$= \ln\frac{V_+e^{-\alpha z}}{V_+e^{-\alpha(z+d)}} = \ln e^{\alpha d} = \alpha d \qquad \text{(Np)} \tag{18.37}$$

The unit of the attenuation constant, $\alpha$, is thus *neper per meter* (Np/m).

The attenuation of voltage or current along a line section is more often expressed in terms of decimal logarithm in *decibels* (dB) (after Alexander Graham Bell, 1847–1922, inventor of the telephone), as

Attenuation of forward voltage wave in decibels

$$= 20\log\frac{V_+e^{-\alpha z}}{V_+e^{-\alpha(z+d)}} = 20\log e^{\alpha d} = (20\log e)\,\alpha d \qquad \text{(dB)}. \tag{18.38}$$

Since $20\log e = 8.686$, the attenuation in decibels is 8.686 times the attenuation in nepers, or $1\,\text{Np} = 8.686\,\text{dB}$.

*Questions and problems:*   Q18.17, P18.28

## 18.5   Basics of Analysis of Transmission Lines in the Time Domain

For various reasons, cables might have, or develop in use, faults along their length. It is useful to know where, so that they can be quickly repaired without pulling the whole cable out. The instrument used today to find faults in cables is called the *time domain reflectometer (TDR)*. Its operating principle is very simple: the instrument sends a voltage step and waits for the reflected signal. If there is a fault in the cable it will be equivalent to some rapid change in cable properties, and part of the voltage step wave will reflect off the discontinuity. As both the transmitted and reflected waves travel at the same velocity, the distance of the fault from the place where the TDR was connected can be calculated exactly. Not only can we learn where the fault is, but the TDR can also tell us something about the nature of the fault.

So far, we have looked at transmission lines only in the frequency domain (we assumed sinusoidal voltages and currents). Now we will look at what happens when a step function (in time) is launched down a transmission line terminated in a load. To analyze the time-domain response, we first replace the entire line with its Thévenin equivalent with respect to the load, as derived in Example 18.8. We thus obtain a simple circuit with a Thévenin generator connected to a load. Transients in such a circuit can next be analyzed by solving a differential equation, or by the Laplace (or Fourier) transform (the two procedures are basically the same). We will use the latter method, where we multiply the Laplace (or Fourier) transform of the reflection coefficient (i.e., the reflection coefficient in complex form) with the transform of a step function and then transform back to the time domain with the inverse Laplace (or Fourier) transform.

**Example 18.11—Reflection from an inductive load.** Let us consider reflection from an inductive load (Fig. 18.11a). The transmission line has a characteristic impedance $Z_0$ and the incident voltage wave is $v_+(t)$. The Thévenin equivalent generator and impedance for this line are $Z_{Th} = Z_0$ and $v_{Th}(t) = 2v_+(t)$ (Fig. 18.11b).

If we now assume that the incident voltage wave is a unit step function starting at $t = 0$, $v_+(t) = 1$, $t > 0$, the Laplace transform is

$$v_+(s) = \frac{1}{s},$$
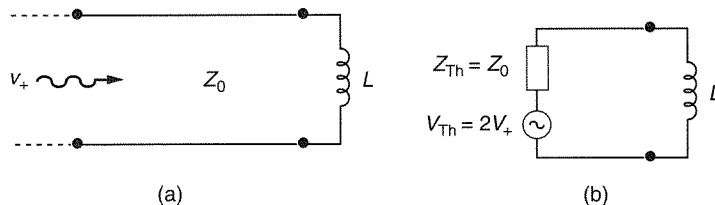


Figure 18.11 (a) A transmission line with an incident wave $v_+$, terminating in an inductive load. (b) The lossless transmission line is replaced by its Thévenin equivalent circuit.

and because the impedance of the inductor is

$$Z = sL.$$

we find that the load voltage is equal to

$$v(s) = \frac{2L}{sL + Z_0} = \frac{2}{s + Z_0/L}. \tag{18.39}$$

We can recognize this as the Laplace transform of a decaying exponential with a time constant $t_L = L/Z_0$:

$$v(t) = 2e^{-t \cdot t_L} \qquad t > 0. \tag{18.40}$$

Because the voltage of an inductor is $v(t) = L \, di/dt$, we can find the current through the inductive load by integrating the voltage:

$$i(t) = \frac{1}{L} \int_0^t v(t) \, dt = \frac{2}{Z_0} (1 - e^{-t/t_L}) \qquad t > 0. \tag{18.41}$$

This describes the buildup of current in an inductor through a resistor, which we already understand from circuit theory. The reflected wave is the difference between the transmitted wave and the incident wave:

$$v_-(t) = v(t) - v_+(t) = 2e^{-t/t_L} - 1 \qquad t > 0. \tag{18.42}$$

We can see that initially the inductor has no current, and the voltage is just $v_+$, so it looks like an open circuit and the reflection coefficient is $+1$. The current then builds up to the short circuit current (the Norton equivalent current) and the voltage drops to zero, so the inductor appears as a short circuit. The reflected and transmitted waves are shown in Fig. 18.12.

**Example 18.12—Reflection from a short circuit.** As another example, let us look at a transmission line that is shorted at one end. If a voltage source is turned on at the other end, what will the reflected wave look like back at the source? We know that the reflection coef-
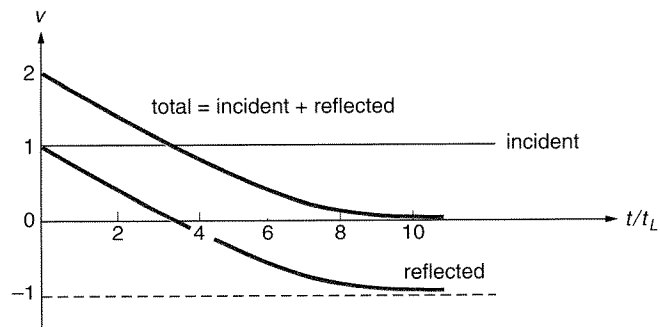


Figure 18.12 The incident, reflected, and total voltages for an inductive load
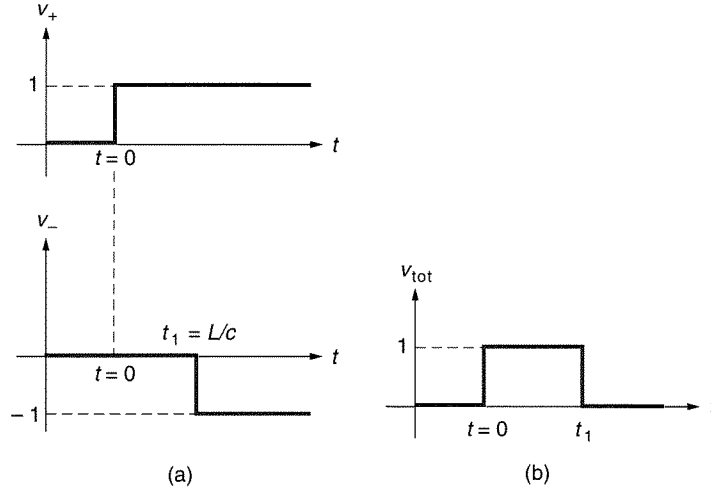
Figure 18.13 (a) Reflected voltage wave off a short-circuited transmission line with an incident unity step function. (b) A standard TDR display shows the reflected wave added on to the incident step function.

ficient of a short circuit is $-1$, so the reflected wave is as shown in Fig. 18.13a. In TDRs, the reflected wave is added on to the incident step (which goes on forever in time), so in this case the instrument display would appear as shown in Fig. 18.13b. The duration of the "pulse" tells us how long the line is (it corresponds to the round-trip time of the leading edge of the step).

**Example 18.13—Reflection from a series RL circuit.** A third, slightly more complicated, example is that of a series combination of an inductor $L$ and a resistor $R$. The incident voltage is a step of unit amplitude. The voltage across the inductor is, as before,

$$v_L(t) = 2e^{-t/t_L}, \quad t > 0, \tag{18.43}$$

where the time constant is now $t_L = L/(R + Z_0)$, because the inductor sees a series connection of the characteristic impedance and the resistive load. The inductor current is

$$i_L(t) = \frac{2}{Z_0 + R}(1 - e^{-t/t_L}), \quad t > 0, \tag{18.44}$$

and the load voltage becomes

$$v(t) = v_L(t) + Ri_L(t) = 2\left[\frac{R}{R + Z_0} + \frac{Z_0}{R + Z_0}e^{-t/t_L}\right], \quad t > 0. \tag{18.45}$$

The reflected voltage wave, shown in Fig. 18.14a, is now

$$v_-(t) = v(t) - v_+(t) = \left[\frac{R - Z_0}{R + Z_0} + 2\frac{Z_0}{R + Z_0}e^{-t/t_L}\right], \quad t > 0. \tag{18.46}$$
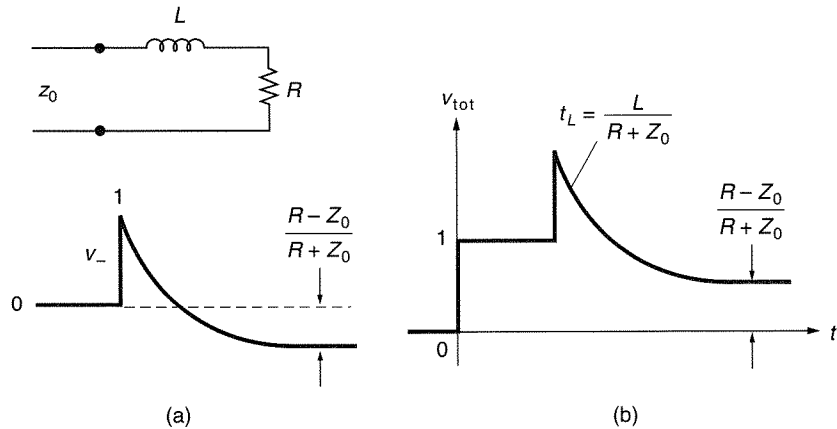
Figure 18.14 (a) Reflected voltage wave off a series $RL$ combination with an incident unity step function. (b) A standard TDR instrument display shows the reflected wave added on to the incident step function.

A simpler qualitative analysis can be done by just evaluating the reflected voltage at $t = 0$ (the time when the reflected wave gets back to the launching port, for example) and $t = \infty$, and assuming any transition between these two values to be exponential. In the previously analyzed case of a series $RL$ circuit, at $t = 0$ the reflected voltage is $v_-(0) = +1$, since the inductor looks like an open circuit initially. On the other hand, as time goes by the current through the inductor builds up, and at $t = \infty$ the inductor looks like a short, so $v_-(\infty) = (R - Z_0)/(R + Z_0)$ and is determined by the resistive part of the load. The resulting plot out of a TDR (incident step plus reflected wave) is shown in Fig. 18.14b.

**Example 18.14—Measuring the time constant of the reflected wave from complex loads.** The most straightforward way to measure the time constant (such as $t_L$ in the inductor examples) is to measure the time $t_1$ needed to complete half of the exponential transition from $v_-(0)$ to $v_-(\infty)$. This corresponds to $t_L = t_1/0.69$, where $t_L$ is the time constant we used for an inductive load, but it also holds for a capacitive load. This procedure is shown qualitatively in Fig. 18.15.

*Questions and problems:*    Q18.18 to Q18.21, P18.29 and P18.30
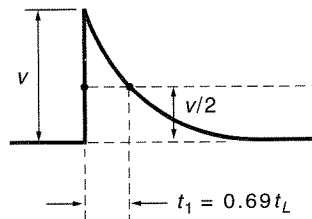


Figure 18.15 Determining the time constant of an exponential TDR response

## 18.6   The Graphical Solution of Lossless-Line Problems Using the Smith Chart

Until the advent of digital computers, the solution of transmission-line problems was done most often with the aid of a graphical tool known as the Smith chart (P. H. Smith, "Transmission-line Calculator," *Electronics*, **12**, January 1939, p. 29; "An Improved Transmission-line Calculator," *Electronics*, **17**, January 1944, p. 130). The Smith chart is a polar plot of the reflection coefficient with some additional details. We restrict our attention to the Smith chart used for solving problems with lossless lines, that is, for $Z_0$ real.

The Smith chart enables us to make a direct determination of the complex reflection coefficient $\rho(0)$ at the load, corresponding to a given load impedance $Z_L$ and the characteristic impedance of the line. Conversely, if $\rho(0)$ is determined experimentally, we can read directly from the chart the load impedance $Z_L$ if $Z_0$ is known. However, the usefulness of the Smith chart far surpasses these two relatively simple tasks, and its use does not seem to decline. Even in the most advanced measurement instruments, such as network analyzers, a Smith chart can be generated on the screen to represent the measurement results, because of the very compact form of such representation. Therefore, we will illustrate the use of a Smith chart with a number of examples. The theoretical basis of the Smith chart is given in most higher-level books on electromagnetics and microwave engineering.

A chart in its usual form, with some additional details whose use will be explained later, is shown in Fig. 18.16. As we have already mentioned, the Smith chart is used for plotting impedances and reflection coefficients. An impedance is plotted on the chart as a *normalized impedance*, defined as

$$z = \frac{Z}{Z_0} = \frac{R + jX}{Z_0} = r + jx \qquad \text{(dimensionless)}. \tag{18.47}$$

*(Definition of normalized load impedance)*

The real part of the impedance, $r$, is defined by the complete circles on the chart. The imaginary part $x$ is defined by the circular arcs. A normalized complex impedance, $z = r + jx$, is defined by the intersection of a circle and an arc. For example, the circle labeled $r = 1$ in Fig. 18.16 intersects the arc labeled $jx = j1$ at the point labeled $z$, which corresponds to an impedance of $Z = z \cdot Z_0 = Z_0(1 + j1)$. If $Z_0 = 50 \ \Omega$, this corresponds to $Z = 50 + j50 \ \Omega$.

The complex reflection coefficient corresponding to $z$ is plotted in polar form, $\rho = |\rho| \angle \phi$, by drawing a straight line from the center of the chart to point $z$. The distance of the point $z$ from the chart center gives $|\rho|$, which can be read off the scale on the horizontal line going through the center of the chart. The angle $\angle \phi$ is read off the (innermost) angular scale on the outer circle of the chart.

The basic properties of the Smith chart are the following:

- All points on the horizontal axis (labeled $r$) correspond to purely real impedances.

Figure 18.16 The Smith chart

- All points on the circle bounding the chart (labeled $x$) correspond to purely imaginary impedances.
- The rightmost point on the chart corresponds to an open circuit (labeled O).
- The leftmost point on the chart corresponds to a short circuit (labeled S).
- The center of the chart has a reflection coefficient equal to zero and corresponds to a matched load (labeled M).
- All points in the lower chart half correspond to loads with a capacitive (negative) reactance.
- All points in the upper chart half correspond to loads with an inductive (positive) reactance.
- The points on the circle labeled $r = 1$ correspond to loads with a real part equal to the adopted normalizing characteristic impedance, $Z_0$ (usually 50 $\Omega$).
- The points on the arc labeled $jx = j1$ correspond to loads with a positive imaginary part equal to the adopted normalizing impedance, $Z_0$ (usually 50 $\Omega$).
- The points on the arc labeled $jx = -j1$ correspond to loads with a negative imaginary part equal to the adopted normalizing impedance, $Z_0$ (usually 50 $\Omega$).
- The length of the straight-line segment between the chart center and a point in the chart represents the magnitude of the reflection coefficient, the points on the boundary circle being of magnitude one. The angle scale on the chart boundary gives the reflection coefficient angle.
- The inside of the Smith chart corresponds to passive impedances (no generators). The magnitude of the reflection coefficient is smaller than or equal to unity inside the chart.
- The outside of the Smith chart (not plotted in Fig. 18.16) corresponds to impedances that give reflection coefficients of magnitudes larger than unity. This means that we can use the chart for active circuits, such as amplifiers and oscillators (i.e., generators). This external part of the chart is sometimes also plotted, and such a chart is referred to as an "extended Smith chart."

**Example 18.15—Determination of the reflection coefficient at the load.** The information that can be obtained directly from a Smith chart is the complex reflection coefficient at the load, $\rho = \rho(0)$, corresponding to a certain normalized load impedance $z = r + jx$. For example, for $Z_0 = 50 \ \Omega$ and $Z_L = (40 + j60) \ \Omega$ we have $z = Z_L/Z_0 = 0.8 + j1.2$, so $r = 0.8$ and $x = +1.2$. From Fig. 18.16, we find that $|\rho| \simeq 0.57$, and $\theta_\rho \simeq 66°$.

The magnitude of $\rho$ is obtained by first measuring the distance of the point $M$ from the chart center (point $\rho' = \rho'' = 0$), using a compass. Below the chart a linear scale is provided, which we can use to obtain the distance measured by the compass in terms of the chart radius (unit circle in the complex $\rho$ plane). For easy reading of the angle $\theta_\rho$, an angle scale marked "angle of reflection coefficient in degrees" is provided around the main chart. So we only need to draw a straight line from the origin through $M$ to determine its intersection with the angle scale.

**Example 18.16—Determination of the load impedance from the reflection coefficient.** The converse problem of determining the normalized load impedance for a given (say, experimentally determined) reflection coefficient at the load is equally simple. For example, according to Fig. 18.16, for $\rho = 0.8 e^{-j45°}$, that is, $|\rho| = 0.8$ and $\theta_\rho = -45°$, we obtain $z \simeq 0.75 - j2.20$. So if, say, $Z_0 = 60\ \Omega$, the load impedance is $Z_L = z \cdot Z_0 \simeq (45.0 - j132)\ \Omega$.

In addition to these two simple applications of the Smith chart, there are several more sophisticated ones. Perhaps the most important is that of determining the input impedance of a line of a given length and characteristic impedance, terminated by a given load impedance. This problem can be solved by means of Eq. (18.28), but an approximate solution using the Smith chart is quite simple. Let us consider the position along the line at a distance $\zeta$ from the load, as in Fig. 18.7. We first normalize $Z(\zeta)$ given in Eq. (18.25) with respect to $Z_0$:

$$z(\zeta) = \frac{Z(\zeta)}{Z_0} = \frac{1 + \rho(0)\,e^{-j2\beta\zeta}}{1 - \rho(0)\,e^{-j2\beta\zeta}}. \tag{18.48}$$

For $\zeta = 0$, $z(\zeta)$ is identical to $z = Z_L/Z_0$. But $z(\zeta)$ is of *exactly* the same form as $z$, except that $\rho(0)$ in $z$ is replaced by $\rho(\zeta) = \rho(0)e^{-j2\beta\zeta}$:

$$z(\zeta) = \frac{1 + \rho(\zeta)}{1 - \rho(\zeta)} \qquad \rho(\zeta) = \rho(0)\,e^{-j2\beta\zeta}. \tag{18.49}$$

Because $|\rho(0)| \le 1$ and $|e^{-2j\beta\zeta}| = 1$, $|\rho(\zeta)| \le 1$. So the chart for $z(\zeta)$ and $\rho(\zeta)$ is exactly the same as for $z$ and $\rho = \rho(0)$. Now, if we know $z$ (for example, $Z_L$ and $Z_0$), we can locate the point on the Smith chart that determines $\rho$ directly. To obtain $z(\zeta)$, however, $\rho(\zeta) = \rho e^{-j2\beta\zeta}$ is needed rather than $\rho$. But multiplying a complex number by $e^{-j2\beta\zeta}$ implies changing its angle by $-2\beta\zeta$, leaving its magnitude constant, which means that we simply have to rotate $\rho$ (corresponding to $z$) by $2\beta\zeta$, in the clockwise (negative) direction. Thus we obtain $\rho(\zeta)$ and can read $z(\zeta)$ directly from the chart.
Noting that

$$2\beta\zeta = 2\frac{2\pi}{\lambda}\zeta = \frac{2\zeta}{\lambda}2\pi, \tag{18.50}$$

it follows that an angle of rotation $2\pi$ corresponds to $\zeta = \lambda/2$. This must be so because we know from Eq. (18.28) that $z(\zeta) = z(\zeta + \lambda/2)$. To facilitate the rotation of $\rho$ by the proper angle, an additional scale around the Smith chart is provided, with 0.5 (wavelengths) corresponding to one complete revolution around the unit circle $|\rho| = 1$. In the chart shown in Fig. 18.16 this wavelength scale is designated by "wavelengths toward generator." For some applications the same scale in the opposite (counterclockwise) direction is also useful and is designated by "wavelengths toward load" in Fig. 18.16.
So we have the following additional properties of the Smith chart related to the reflection coefficient $\rho(\zeta)$:

- Moving around the chart in the clockwise direction corresponds to moving down the line from the load toward the generator (the phase increases).

- Moving around the chart in the counterclockwise direction corresponds to moving down the line from the generator toward the load (the phase decreases).
- One full circle around the chart corresponds to 180 degrees of phase (or half a wavelength). This is because the phase of the reflection coefficient changes as $e^{2j\beta z}$, so everything repeats every half wavelength down a line.

**Example 18.17—Input impedance of a line terminated in an arbitrary impedance.** As an example, let us consider a line of characteristic impedance $Z_0 = 60\ \Omega$ and of length $\zeta = 0.40\,\lambda$ at the frequency used. Let us suppose that $Z_L = (90 - j60)\ \Omega$, and that we wish to determine the input impedance of the line thus terminated, using the Smith chart.

First, $z = Z_L/Z_0 = 1.5 - j1$, and we start with this value in the chart. This point has to be rotated in a clockwise direction by 0.4 units on the wavelength scale. Therefore we draw a straight line from the center of the chart through the point $z = 1.5 - j1$. The intersection of this line and the "wavelengths toward generator" scale is at 0.308 on the scale. We add 0.40 to this and get 0.708. This is 0.208 farther than point 0.000 on the scale. We draw a straight line from the 0.208 point of the wavelength scale toward the chart center and measure along this line the distance of the point $z = 1.5 - j1$ from the center. The point found in this way determines $z(\zeta) = z(0.4\lambda)$. From the chart we find that $z(0.4\lambda) \simeq (1.83 + j0.95)$. So the input impedance of a $0.4\lambda$ long 60-$\Omega$ line terminated with $Z_L = (90 - j60)\ \Omega$ is $Z(0.4\lambda) = Z_0 z(0.4\lambda) \simeq (110 + j57.0)\ \Omega$.

**Example 18.18—Examples of matching by transmission-line segments.** As already mentioned, at high frequencies (above about 100 MHz) it is not simple to make passive elements like resistors, capacitors, inductors, and transformers. For example, shunt (parallel) susceptance of interwinding capacitances of coils at these frequencies becomes pronounced and may completely distort the frequency behavior of the inductor. Shorted or open sections of transmission lines do not have this problem, so they are frequently used to replace reactive circuit elements in such cases. Such transmission-line reactive elements are often used for matching a high-frequency load to a desired impedance.

Another possible use of transmission-line segments for matching is as components for matching a load to a transmission line of a given characteristic impedance. Three principal ways of using transmission-line matching sections are sketched in Fig. 18.17.
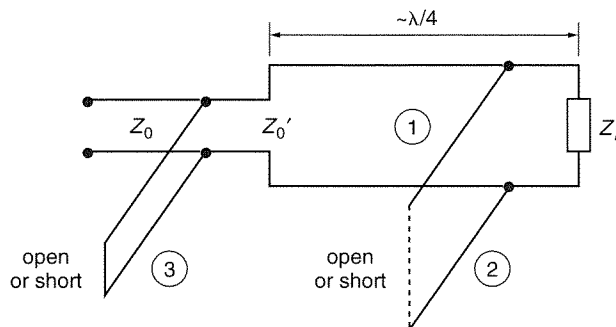


Figure 18.17 Three principal ways of using transmission-line sections for matching purposes: (1) a matching line section; (2) stub matching at the load; (3) stub matching along the line

Suppose the load impedance is $Z_L = R_L + jX_L$ and the transmission-line characteristic impedance is $Z_0 \neq R_L$. We can attempt to match the load to the line by following two steps:

1. Add a shorted line section in parallel to the load (a "stub" labeled 2 in Fig. 18.17), such that the admittance (and impedance) of the combination becomes real. Let the impedance of the combination be $Z'_L$.

2. Add a quarter-wavelength *matching line section* like that labeled 1 in Fig. 18.17 to match the load $Z'_L$ to $Z_0$.

In some instances, a matching line section of length different than quarter wavelength may do the entire job when it transforms the load impedance to approximately $Z_0$ without the stub at the load.

Finally, it is possible to add another stub, labeled 3 in Fig. 18.17, at a convenient location along the line to improve matching. Although all such problems can be solved with ease by programmable calculators or computers, they can also be solved simply using the Smith chart. Several specific examples of matching are given in the problems at the end of the chapter.

These examples illustrate only some of the simplest applications of the Smith chart. Several others will be found in the problems at the end of the chapter. Applications of the Smith chart are much more diverse than these examples suggest. The chart can also be used for analyzing plane waves perpendicularly incident on a plane boundary surface, for analyzing lossy lines, and for many other problems. Even though we can use a computer to perform such tasks, the Smith chart is useful for presenting the results and getting an intuitive feel for what the analysis tells us.

*Questions and problems:*   Q18.22 and Q18.23, P18.31 to P18.40

## 18.7   Chapter Summary

1. A transmission line is an electromagnetic structure, so strictly speaking it should be analyzed by means of the field equations. However, a very short segment of a line can be approximated by a simple circuit, and the complete line by a chain of such circuits. Consequently, transmission lines can also be analyzed using circuit theory.

2. The voltage and current along transmission lines have a property not encountered in "normal" circuits: they move along the line with a certain velocity. These moving voltages and currents are known as voltage and current waves.

3. If the line is infinitely long, the ratio of the voltage and current waves propagating in one direction along the line, at any point and at any instant, is constant. This constant is known as the line characteristic impedance, and it depends only on how the line is made (dimensions and materials).

4. For lossless air lines, the velocity of propagation of voltage and current waves along them equals the velocity of light in a vacuum, whereas in all other cases this velocity is smaller.

5. The input impedance of open- or short-circuited transmission-line segments is purely reactive. Therefore such segments are used at high frequencies as capacitors and inductors.

6. Transmission-line segments act as specific, frequency-dependent transformers of impedances of loads connected at their end. This, combined with adding appropriate segments (stubs) of shorted (or open) lines in parallel, can be used for matching a transmission line to the load it is terminated in.

# QUESTIONS

**Q18.1.** Why is it not practically possible to obtain a coaxial cable of characteristic impedance $Z_0 = 500 \; \Omega$? Can you have a two-wire line of this characteristic impedance?

**Q18.2.** Assume that a transmission line is made of two parallel, highly resistive wires. Can this line be analyzed using fundamental transmission-line equations? Explain.

**Q18.3.** A coaxial cable is filled with water. Does it represent a transmission line? Explain.

**Q18.4.** Two wires several wavelengths long serve as a connection between a generator and a receiver. The distance between the wires is small but not constant, varying as a smooth function of the coordinate along the line. Can you use the transmission-line equations for the analysis of this line? Explain.

**Q18.5.** Explain how you can obtain (1) a forward wave only; (2) a backward wave only along a transmission line.

**Q18.6.** Describe at least three ways of obtaining simultaneously a forward and a backward sinusoidal wave of the same amplitude along a transmission line.

**Q18.7.** Why can you replace an infinitely long end of a transmission line with a resistor of resistance equal to the line characteristic impedance?

**Q18.8.** Can a voltage (or a current) wave along a transmission line be described by the expression of the form $u(xy)$, where $u(xy)$ is a function of the product of the arguments $x = (t - z/c)$ and $y = (t + z/c)$? Explain.

**Q18.9.** Can we adopt the negative instead of positive value of the square root in Eq. (18.7) for the velocity of wave propagation along transmission lines? Explain.

**Q18.10.** Why must the exponent of the forward voltage and current waves in Eqs. (18.13) and (18.17) be negative? Why must those of the backward waves be positive?

**Q18.11.** Is the wavelength along an air line greater or less than that in the same line filled with a dielectric? What is the answer if the dielectric has relative permeability greater than one? Explain.

**Q18.12.** What are the SI units for the following quantities: (1) the attenuation constant ($\alpha$), (2) the phase constant ($\beta$), (3) the reflection coefficient ($\rho$), and (4) the voltage standing-wave ratio (VSWR)?

**Q18.13.** What is the magnitude of the reflection coefficient, $|\rho|$, and of the VSWR, for which one half of the power of the incident wave is transferred to the load?

**Q18.14.** Why is the voltage at the termination $Z$ of a transmission line with characteristic impedance $Z_0$ equal to $V = 2V_+ Z/(Z + Z_0)$?

**Q18.15.** What are the input impedances to lossless lines of lengths $\lambda/4$ and $\lambda/2$, if they are (1) open-circuited or (2) short-circuited?

**Q18.16.** Can a resistive load of *any* resistance $R$ be matched in practice to a transmission line of characteristic impedance $Z_0$? Explain.

**Q18.17.** The characteristic impedance of a lossy line in Eq. (18.32) is real if $R' = 0$ and $G' = 0$. Can it be real for some other relation between $R'$, $L'$, $G'$, and $C'$? Explain.

**Q18.18.** Why could we not use simple transmission-line analysis when calculating the step response of an inductor, as in Fig. Q18.18?
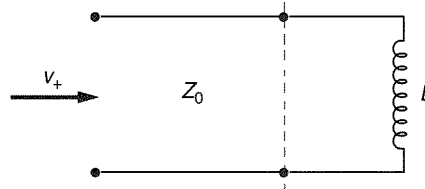


Figure Q18.18 Calculating the step response of an inductor

**Q18.19.** If you had a break in the dielectric of a cable causing a large shunt conductance, what do you expect to see reflected if you excite the cable with a short pulse (practical delta function)?

**Q18.20.** If you had a break in the outer conductor of a cable, causing a large series resistance, what do you expect to see reflected if you excite the cable with a short pulse (imperfect delta function)?

**Q18.21.** What do the reflected waves off a series inductor and shunt capacitor in the middle of a transmission line look like for a short pulse excitation, assuming that $\omega L \gg Z_0$ and $\omega C \gg 1/Z_0$?

**Q18.22.** Using the Smith chart, determine the complex reflection coefficient on a 60-$\Omega$ line if it is terminated by (1) 80 $\Omega$, (2) $(30 - j40)$ $\Omega$, or (3) $(40 + j90)$ $\Omega$.

**Q18.23.** Using the Smith chart, determine the terminating impedance of a 70-$\Omega$ line if it was found experimentally that the complex voltage reflection coefficient is (1) 0.8, (2) $0.2e^{-j\pi/4}$, or (3) $0.5e^{j\pi/3}$.

## Problems

**P18.1.** Given a high-frequency RG-55/U coaxial cable with $a = 0.5$ mm, $b = 2.95$ mm, $\epsilon_r = 2.25$ (polyethylene), and $\mu_r = 1$, find the values for the capacitance and inductance per unit length of the cable.

**P18.2.** Assume that the coaxial cable from problem P18.1 is not lossless but that the losses are small, resulting in an attenuation constant in decibels per meter at 10 GHz of $\alpha = 0.5$ dB/m. Assuming the dielectric in the cable to be perfect, find the resistance per unit length that causes the losses in the conductors.

**P18.3.** The distance $d$ between wires of a lossless two-wire line is a smooth, slowly varying function of the coordinate $z$ along the line so that the line capacitance and inductance per unit length, $L'$ and $C'$, are also smooth functions of $z$, $L' = L'(z)$, and $C' = C'(z)$. Derive the transmission-line equations for such a nonuniform transmission line. Check

if these equations become the transmission-line equations (18.4) for $L'(z)$ and $C'(z)$ constant.

**P18.4.** Using circuit theory, analyze approximately a matched, lossless, air-filled coaxial transmission line of length $l = \lambda$ and conductor radii $a = 1$ mm and $b = 3$ mm as a connection of $n$ cells of the type in Fig. 18.3b, for $n = 1, 2, \ldots, 20$. Such a circuit-theory approximation to transmission lines is known as an *artificial transmission line*. Note that an artificial transmission line can be analyzed as a simple ladder network. Assume the artificial line to be terminated in the actual characteristic impedance, and compare current in series-concentrated inductive elements and voltage across parallel concentrated capacitive elements with exact results. Solve the problem so that you can vary $L'$, $C'$, and $n$.

**P18.5.** Noting that $c = 1/\sqrt{\epsilon\mu}$ for all transmission lines in Table 18.1, prove that for these lines the inductance per unit length and the characteristic impedance of a lossless transmission line can be expressed in terms of $c$ and $C'$.

**P18.6.** Express $V(z)$ in Eq. (18.13) and $I(z)$ in Eq. (18.17) for lossless lines in terms of the sending-end voltage $V(0)$ and sending-end current $I(0)$.

**P18.7.** Prove that it is possible to obtain the characteristic impedance of any lossless line by measuring the input impedance of a section of the line when it is open-circuited, and when it is short-circuited.

**P18.8.** A lossless line of characteristic impedance $Z_{01}$ and length $l_1$ is terminated in an impedance $Z_L$. The line serves as a load for another lossless line of characteristic impedance $Z_{02}$ and length $l_2$. The dielectric in both lines is air and the angular frequency of the current is $\omega$. Determine general expressions for the input impedance of the second line, the reflection coefficient in both lines, and the voltage standing-wave ratio in both lines.

**P18.9.** A short and then an open load are connected to a 50-$\Omega$ transmission line at $z = 0$. Make a plot of the impedance, normalized voltage ("normalized" means that you divide the voltage by its maximal value to get a maximum normalized voltage of 1), and normalized current along the line up to $z = -3\lambda/2$ for the two cases.

**P18.10.** A *lumped* capacitor is inserted into a transmission-line section, as shown in Fig. P18.10. Find the reflection coefficient for a wave incident from the left. Assume the line is terminated to the right so that there is no reflection off the end of the line. Find a simplified expression that applies when $C$ is small. The characteristic impedance of the line is $Z_0$.
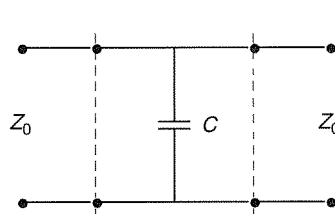


Figure P18.10 A shunt capacitor in a line
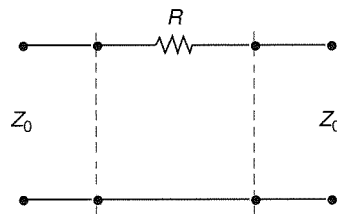
Figure P18.11 A series resistor in a line

**P18.11.** Repeat problem P18.10 assuming that a lumped resistor is inserted into a transmission-line section as shown in Fig. P18.11. Find a simplified expression that applies when $R$ is small.

**P18.12.** Repeat problem P18.10 assuming that a lumped inductor is inserted into a transmission-line section as shown in Fig. P18.12. Find a simplified expression that applies when $L$ is small.
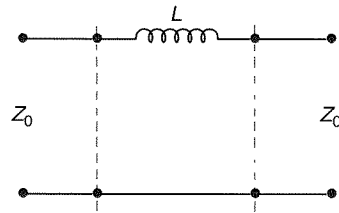


Figure P18.12 A series coil in a line
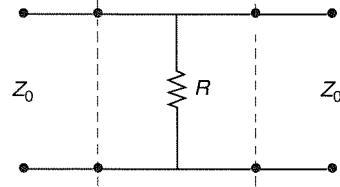
Figure P18.13 A shunt resistor in a line

**P18.13.** Repeat problem P18.10 assuming that a lumped resistor is inserted into a transmission-line section as shown in Fig. P18.13. Find a simplified expression that applies when $R$ is large.

**P18.14.** A 50-$\Omega$ transmission line needs to be connected to a 100-$\Omega$ load. The setup is used at 1 GHz. What would you connect between the line and the load to have no reflected voltage on the line?    * LOSSLESS ELEMENTS

**P18.15.** In problem P18.14, the load is a 100-$\Omega$ resistor but the leads are long and represent a 2 nH inductor in series with the resistor. How would you get rid of the reflected voltage on the line in this case?

**P18.16.** Find the transmission coefficient for the transmission line in Fig. P18.10.

**P18.17.** Find the reflection and transmission coefficients for the transmission line in Fig. P18.17. Because the reflection coefficient is defined by voltage, the power is given by its square. What are the reflected and transmitted power equal to? Does the power balance make sense?
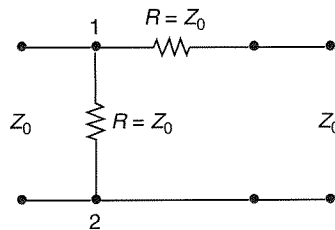


Figure P18.17 Two resistors in a line

**P18.18.** Derive the normalized input impedance (i.e., the impedance divided by $Z_0$) for a section of line that is $n\lambda/8$ long and shorted at the other end, for $n = 1, 2, 3, 4$, and 5. Plot the impedance as a function of electrical line length from the load (length measured in wavelengths along the line).

**P18.19.** Repeat problem P18.18 for an open-ended line.

**P18.20.** Find the total current and voltage at the beginning of a $\lambda/4$ shorted transmission line of characteristic impedance $Z_0$. What circuit element does this line look like? Plot the

total current and voltage as a function of electrical line length from the load (length measured in wavelengths along the line).

**P18.21.** Repeat the previous problem for an open-ended line.

**P18.22.** Find the reflection and transmission coefficients for an ideal $n : 1$ transformer, as in Fig. P18.22, where $n$ is the voltage transformation ratio.
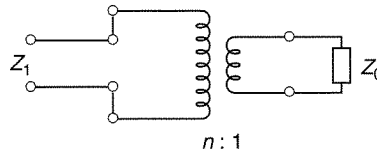


Figure P18.22 An ideal transformer

**P18.23.** Find the input impedance for the circuit in Fig. P18.23.
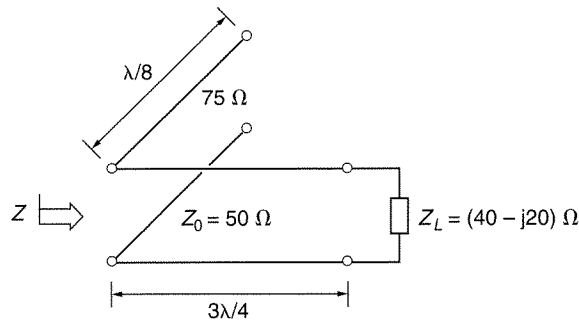


Figure P18.23 Impedance of a line with a shunt stub

**P18.24.** A coaxial transmission line with a characteristic impedance of 150 $\Omega$ is 2 cm long and is terminated in a load impedance of $Z = 75 + j150$ $\Omega$. The dielectric in the line has a relative permittivity of $\epsilon_r = 2.56$. Find the input impedance and VSWR on the line at $f = 3$ GHz.

**P18.25.** Match a 25-$\Omega$ load to a 50-$\Omega$ line using (1) a single quarter-wave section of line, or (2) two quarter-wave line sections.

**P18.26.** Match a purely capacitive load, $C = 10$ pF, to a 50-$\Omega$ line at 1 GHz. How many different ways can you think of doing this?

**P18.27.** Calculate and plot magnitude and phase of $\rho(f)$ between 1 and 3 GHz for a 50-$\Omega$ open transmission line that is $\lambda/4$ long at 2 GHz.

**P18.28.** If you had a cable like the one in problem P18.2 spanning the Atlantic and you sent a continuous signal of 1-MW power from the United States to England, how much power approximately would you get in England? (Look up the approximate distance across the Atlantic in an atlas if you need to.)

**P18.29.** A printed-circuit board trace in a digital circuit is excited by a voltage $v(t)$, as in Fig. P18.29. Derive an equation for the coupled (cross-talk) signal on an adjacent line, $v_c(t)$, assuming the adjacent line is connected to a load at one end and a scope (infinite

impedance) at the other end so that no current flows through it. (*Hint:* the coupling is capacitive and you can approximate it by a capacitor between the two traces and use circuit theory.)
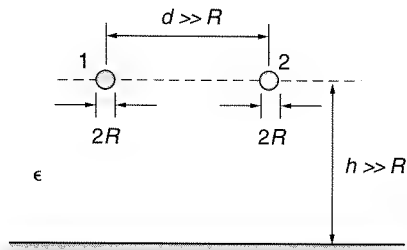


Figure P18.29 An example of two coupled lines
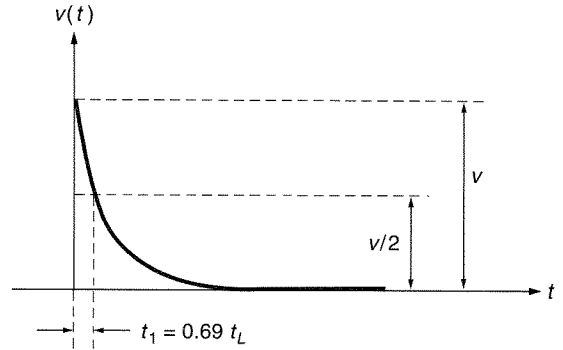


Figure P18.30 Measuring the reflected wave from a complex load

**P18.30.** Derive the expression $t_1 = 0.69 \, t_L$ discussed in Example 18.14. This expression shows a practical way to measure the time constant of the reflected wave for the case of complex loads, as in Fig. P18.30.

**P18.31.** Trace the procedure for solving problem P18.8 by means of the Smith chart.

**P18.32.** The reciprocals of complex numbers can be determined easily from the Smith chart. Starting with Eq. (18.49), deduce how this can be done.

**P18.33.** A fixed, known complex impedance $Z_L$ is to be connected to a lossless line having a characteristic impedance $Z_0$. Show that it is possible to eliminate the reflected wave along the line if an appropriate length of the same line, assumed to be short-circuited, is connected at an appropriate place on the line near $Z_L$ (see Fig. P18.33).
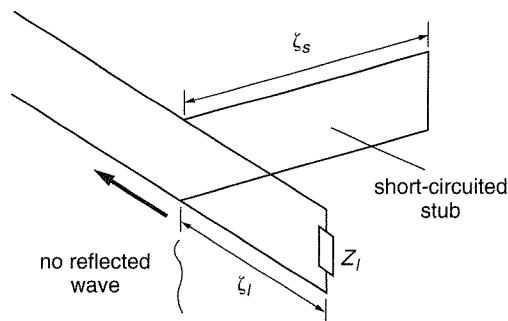


Figure P18.33 A configuration for matching a load to a transmission line

**P18.34.** What circuit element corresponds to the point on the Smith chart that is defined by the intersection of the circle $r = 1$ and the arc $jx = j1.2$ at 1 GHz, if the normalizing impedance is 50 Ω?

**P18.35.** What circuit element corresponds to the point on the Smith chart that is defined by the intersection of the circle $r = 1$ and the arc $jx = -j0.4$ at 500 MHz, if the normalizing impedance is 50 Ω?

**P18.36.** At the load of a terminated transmission line of characteristic impedance $Z_0 = 100$ Ω, the reflection coefficient is $\rho = 0.56 + j0.215$. What is the load impedance?

**P18.37.** A 50-Ω line is terminated in a load impedance of $Z = 80 - j40$ Ω. Find the reflection coefficient of the load and the VSWR.

**P18.38.** A 50-Ω slotted line measurement (see Example 18.10) was done by first placing a short at the place of the unknown load. This results in a large VSWR on the line with sharply defined voltage minima. On an arbitrarily positioned scale along the air-filled coaxial line, the voltage minima are observed at $z_s = 0.1, 1.1,$ and 2.1 cm. The short is then replaced by the unknown load, the VSWR is measured to be 2, and the voltage minima (not as sharp as with the short termination) are found at $z = 0.61, 1.61,$ and 2.61 cm. Use the Smith chart to find the complex impedance of the load. Explain all your steps.

**P18.39.** Use a shorted parallel stub to match a 200-Ω load to a 50-Ω transmission line. Include a Smith chart plot with step-by-step explanations.

**P18.40.** A load consists of a 100-Ω resistor in series with a 10-nH inductor at 1 GHz. Use an open single stub to match it to a 50-Ω line.

# 19

# Maxwell's Equations

## 19.1 Introduction

This chapter is devoted to the extension of the equations we have derived so far to the most general equations for the electromagnetic field. These general equations are known as *Maxwell's equations*. Any engineering problem that includes electromagnetic fields is solved starting from these equations, although in some instances the application may not be obvious. For example, ac circuits are in fact described by an approximation of Maxwell's equations valid for specific fields existing in such circuits.

We will see that Maxwell's equations can be written in two forms: integral and differential. We will also see that numerous general conclusions follow from these equations. For example, the problem of energy transfer by means of an electromagnetic field can be understood and solved only if we start from Maxwell's equations and derive what is known as *Poynting's theorem*. General boundary conditions will also be derived. Finally, we will show that in many important instances electromagnetic field vectors can be derived from auxiliary functions, known as *potentials*.

This is probably the most important chapter in the entire book. It unifies all the concepts we have studied so far. It also adds the concept of displacement current that couples Gauss', Ampère's, and Faraday's laws with the current continuity and conservation of magnetic flux equations. Maxwell's equations enable us to solve many practical engineering problems that deal with electromagnetic fields.

## 19.2 Displacement Current

We know from Faraday's law that a time-varying magnetic field is always accompanied by a time-varying electric (induced) field. This also means that a time-varying electric field is accompanied by a time-varying magnetic field. We have learned so far that sources of a magnetic field are electric currents. From the preceding inverse statement, we can say that a *time-varying magnetic field* is not caused solely by time-varying electric currents but also by a *time-varying electric field*.

This conclusion is the essence of Maxwell's contribution to the theory of electricity and magnetism. To stress that this time-varying electric field is the source of a magnetic field, as is a current, a quantity tightly connected with time variation of the electric field is termed the *displacement current*, even though it is not a current in the usual sense.

Consider a circuit containing an air-filled parallel-plate capacitor and with time-varying current flowing through it, as sketched in Fig. 19.1. Imagine two surfaces, $S_1$ and $S_2$, shown in the figure. The surface $S_1$ intersects a part of the wire. The surface $S_2$ intersects *one electrode of the capacitor only*.

If we apply the current continuity equation [Eq. (10.14)],

$$\int_S \mathbf{J} \cdot d\mathbf{S} = -\frac{d}{dt} \int_v \rho \, dv, \tag{10.14}$$

to $S_1$, we find that it is satisfied because a current $i(t)$ enters the surface, the same current leaves the surface, and there is no charge accumulation along the enclosed wire segment. If, however, we apply Eq. (10.14) to $S_2$, we are working with a current entering $S_2$, but no current leaving this surface. Instead, we have an increase of charge in $S_2$ such that Eq. (10.14) is satisfied.

Suppose we wish to express the general equation for current continuity in Eq. (10.14) as a surface integral on the left-hand side, and a zero on the right-hand side. This can be done easily if we recall Gauss' law in Eq. (7.20). The volume integral
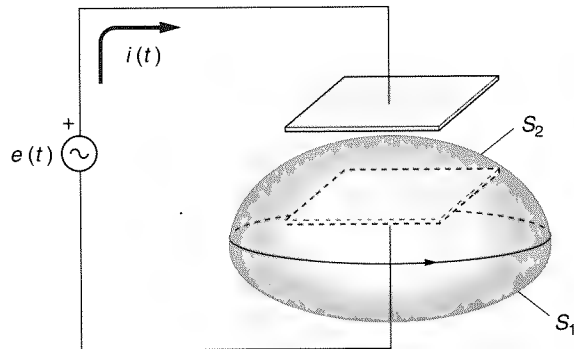


Figure 19.1 Circuit containing an air-filled capacitor and with time-varying current flowing through it

in Eq. (10.14) is precisely $Q_{\text{free in } S}$ in Eq. (7.20), except that this charge now varies in time. So instead of Eq. (10.14) we can write an equivalent equation

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = -\frac{d}{dt} \oint_S \mathbf{D} \cdot d\mathbf{S}. \tag{19.1}$$

If we assume that the surface $S$ is not varying in time, the time derivative can be introduced under the integral sign to act on the vector $\mathbf{D}$ only. Noting that the surface integrals on the two sides of the equation refer to the same surface, we can write Eq. (19.1) in the form

$$\oint_S \left( \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot d\mathbf{S} = 0. \tag{19.2}$$

We have arrived at an interesting conclusion: the flux through a closed surface of the vector sum $(\mathbf{J} + \partial \mathbf{D}/\partial t)$ is always zero. The expression $\partial \mathbf{D}/\partial t$ has the dimension of current density. It is therefore termed the *displacement current density*.

We know that if the flux of a vector function through any closed surface is zero, then the flux of that vector through all *open* surfaces bounded by the same contour is the same. Consider the contour $C$ indicated in Fig. 19.1, and two surfaces bounded by the contour, $S_1$ and $S_2$. The surface $S_1$ cuts the wire, so the flux of $(\mathbf{J} + \partial \mathbf{D}/\partial t)$ through it is simply $i(t)$. The surface $S_2$ passes between the capacitor electrodes and does not cut the wire. Therefore, there is no current through that surface, and the flux of $(\mathbf{J} + \partial \mathbf{D}/\partial t)$ equals that of vector $\partial \mathbf{D}/\partial t$ through it. We will now show that these two integrals are equal.

Open surfaces $S_1$ and $S_2$ make the *closed* surface $S$. The flux of $(\mathbf{J} + \partial \mathbf{D}/\partial t)$ through $S$ is calculated with respect to the *outward* unit vector normal to $S$. Recall the right-hand rule of defining a unit vector normal to a surface defined by a contour. The flux through the part $S_1$ of $S$ is calculated with respect to the outward normal, but the flux through the part $S_2$ of $S$ should be calculated with respect to the opposite normal. Consequently Eq. (19.2) yields

$$\int_{S_1} \mathbf{J} \cdot d\mathbf{S} = \int_{S_2} \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{S}.$$

This could be interpreted as if the conductive current in the metallic wire continues between the capacitor plates in the form of the displacement current. In other words, if we consider a time-varying conductive current only, it has sources and sinks. The *total current* (the sum of conductive and displacement currents), however, does not have sources and sinks, but rather closes onto itself, as a dc current. With this in mind, Maxwell postulated that in time-varying fields the source of the magnetic field is not solely the conductive current, but rather the total current, the density of which is $(\mathbf{J} + \partial \mathbf{D}/\partial t)$.

From Ampère's law we know that the line integral of the magnetic field intensity vector, $\mathbf{H}$, along a closed contour equals the current through any surface defined by the contour. From the reasoning we just did, we see that it is also equal to the flux of the *displacement current* through the contour (i.e., through a surface bounded by

the contour):

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_{S_1} \mathbf{J} \cdot d\mathbf{S} = \int_{S_2} \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{S}. \tag{19.3}$$

This equation tells us that if we wish Ampère's law to be valid for time-varying currents, we must replace $\mathbf{J}$ in it by $(\mathbf{J} + \partial \mathbf{D}/\partial t)$. So the generalized Ampère's law has the form

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \left( \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot d\mathbf{S} \qquad \text{(generalized Ampère's law).} \tag{19.4}$$

This is the fundamental contribution of Maxwell, which can be interpreted as follows: *the displacement current produces a magnetic field according to the same law as "normal" current.* We will see that the addition of the displacement current density in Ampère's law has far-reaching consequences. For example, without it we cannot explain the existence of electromagnetic waves. An electromagnetic wave is a moving electromagnetic field that, once created by charges and currents, continues to exist with no connection whatsoever to the charges and currents that created it.

**Example 19.1—Displacement current density in dielectrics and in a vacuum.** Since $\mathbf{D} = (\epsilon_0 \mathbf{E} + \mathbf{P})$, the displacement current density, $\partial \mathbf{D}/\partial t$, can be written in the form

$$\frac{\partial \mathbf{D}}{\partial t} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \frac{\partial \mathbf{P}}{\partial t}.$$

We know that the polarization vector, $\mathbf{P}$, represents the transfer of *real* charge per unit area normal to vector $\mathbf{P}$. It is measured in $C/m^2$. Therefore the expression $\partial \mathbf{P}/\partial t$ is in $A/m^2$, and represents a *real* current density, resulting from the motion of the polarization charges. This part of the displacement current density is termed the *displacement current density in the dielectric*, or frequently, the *density of polarization current*.

The other part of the displacement current density, $\epsilon_0 \, \partial \mathbf{E}/\partial t$, is measured in the same units, $A/m^2$, but it does not represent any motion of real charges. This is the *displacement current density in a vacuum*. This part of the displacement current can be very misleading, however, if one does not keep in mind its physical meaning: the time-varying electric field is the source of the time-varying magnetic field. In other words, as far as the source of the magnetic field is concerned, $\epsilon_0 \, \partial \mathbf{E}/\partial t$ is completely equivalent to an electric current of the same density, although it does *not* represent any real motion of electric charges.

*Questions and problems:*   Q19.1, P19.1 to P19.3

## 19.3   Maxwell's Equations in Integral Form

We are now ready to formulate the general equations of the electromagnetic field in integral form. In fact, what we need to do is to review all the equations we have postulated or derived, and make sure they are not contradictory. If they do not contradict each other, we can, following Maxwell, *postulate* them to be true for all electromagnetic fields. The sole criterion for the validity of these equations is, of course, experiment. Ever since Maxwell postulated in the 1860s the equations that bear his name,

no experimental evidence has indicated even the slightest disagreement with these equations.

We now write the integral form of the four most general equations we have derived. With no particular reason except that it is customary, we start with Faraday's law in Eq. (14.6) for a fixed contour, so that the time derivative can be introduced under the integral sign. This equation is usually followed by the generalized Ampère's law. Gauss' law in Eq. (7.20), in which the total free charge enclosed by a closed surface is replaced by a volume integral of the charge density, is the third equation. The last equation is the law of conservation of magnetic flux.

Thus Maxwell's equations in integral form are

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S}, \tag{19.5}$$

*[Faraday's law for a fixed contour, Eq. (14.6) = Maxwell's first equation]*

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \left( \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot d\mathbf{S}, \tag{19.6}$$

*[Generalized Ampère's law, Eq. (19.4) = Maxwell's second equation]*

$$\oint_S \mathbf{D} \cdot d\mathbf{S} = \int_v \rho \, dv, \tag{19.7}$$

*[Gauss' law, Eq. (7.20) = Maxwell's third equation]*

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0. \tag{19.8}$$

*[Law of conservation of magnetic flux, Eq. (12.11) = Maxwell's fourth equation]*

Finally, we add to these equations the current continuity equation,

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = -\int_v \frac{\partial \rho}{\partial t} dv. \tag{19.9}$$

*[Current continuity equation, Eq. (10.15) = law of conservation of electric charge]*

These equations can be paraphrased as follows. Equation (19.5) tells us that a time-varying magnetic field is a source of an (evidently time-varying) electric field. Equation (19.6) states that the sources of a magnetic field are electric currents *and a time-varying electric field*. According to Eq. (19.7), the only source that produces a nonzero flux through a closed surface of the electric displacement vector are free electric charges. Finally, Eq. (19.8) can be interpreted as stating that no analogue of free electric charges exists for a magnetic field. Equation (19.9) is not a field equation, but the law of conservation of electric charge must be satisfied by all real sources of the electromagnetic field.

If we try to find any logical deficiencies in these equations, we will see that there are none, in spite of the fact that they have been derived separately, for specific types of fields. For this reason we *postulate* that these equations are always valid and represent the equations of the general electromagnetic field.

There are numerous applications of Maxwell's equations in integral form. One group of applications relates to the derivation of some general conclusions about electromagnetic fields. One of the most important applications of this type is the derivation of general boundary conditions.

**Example 19.2—General boundary conditions.** We know that boundary conditions are relations between values of any field quantity at two close points on the two sides of a surface between two different media. For the four basic field vectors, $\mathbf{E}$, $\mathbf{H}$, $\mathbf{D}$, and $\mathbf{B}$, they are but special forms of the integral Maxwell's equations (19.5) to (19.8).

In order to be able to derive them in the most usual form, we need to consider also the possibility of surface currents. We shall see in the next chapter that at high frequencies, currents in good conductors are distributed essentially over conductor surfaces, and are practically surface currents. This is why we need to include surface currents in boundary conditions, and to specialize the boundary conditions at the surface of a "perfect" conductor.

If one of the two media on two sides of a boundary surface is a perfect conductor, let it be medium 2. Inside a perfect conductor *there can be no electric field* (it would result in infinite current density). We know that a time-varying electric field is accompanied by a time-varying magnetic field. Therefore, inside a perfect conductor, *there can also be no time-varying magnetic field*.

We derived boundary conditions in the electrostatic field starting, in fact, from Eq. (19.5) (with zero right-hand side), and from Eq. (19.7). Does the nonzero right-hand side in Eq. (19.5) change anything? Recall that in the derivation of the boundary condition for the tangential components of vector $\mathbf{E}$ we assumed that the contour was infinitely narrow. Therefore, the flux of vector $\partial \mathbf{B}/\partial t$ is zero also if we start from Eq. (19.5). On the other hand, Eq. (19.7) is the same as Gauss' law in electrostatics. So we conclude that in *any* electromagnetic field, on the two sides of *any* boundary surface, both electrostatic conditions, Eqs. (7.26) and (7.27), remain valid. If one of the media is a perfect conductor, these equations take the forms that are also valid in electrostatics (but for any, not necessarily perfect, conductor):

$$\mathbf{E}_{1\,\text{tang}} = \mathbf{E}_{2\,\text{tang}}, \quad \text{or} \quad \mathbf{E}_{\text{tang}} = 0 \quad \text{on surface of perfect conductor,} \qquad (19.10)$$

*(General boundary condition for tangential components of $\mathbf{E}$)*

and

$$\mathbf{D}_{1\,\text{norm}} - \mathbf{D}_{2\,\text{norm}} = \sigma, \quad \text{or} \quad \mathbf{D}_{\text{norm}} = \sigma \quad \text{on surface of perfect conductor. (19.11)}$$

*(General boundary condition for normal components of $\mathbf{D}$)*

The condition for the tangential components of the magnetic field intensity vector was derived from Ampère's law, applied to an infinitely narrow contour. Displacement current through such a contour is zero, but conduction current may be nonzero if there is a surface current on the boundary.

Consider Fig. 19.2 and assume the surface-current density vector $\mathbf{J}_s$ to be locally in the $y$ direction. The magnetic field of these currents is then in the $x$ direction, as indicated. The current through the narrow rectangular contour in the figure, which is in the $x$-$z$ plane, i.e., normal to $\mathbf{J}_s$, is $J_s\,\Delta l$. The integral of vector $\mathbf{H}$ around the contour is $(H_{1x} - H_{2x})\Delta l$. Noting that the unit vector normal to the boundary is directed into medium 1, from the integral form of Ampère's law we obtain

$$\mathbf{H}_{1\,\text{tang}} - \mathbf{H}_{2\,\text{tang}} = \mathbf{J}_s \times \mathbf{n}, \quad \text{or} \quad \mathbf{H}_{\text{tang}} = \mathbf{J}_s \times \mathbf{n} \quad \text{on surface of perfect conductor. (19.12)}$$

*(General boundary condition for tangential components of $\mathbf{H}$—see Fig. 19.2)*

The condition for the normal components of vector $\mathbf{B}$, Eq. (13.8), also remains the same, since it was derived from the law of conservation of magnetic flux, Eq. (19.8). If medium 2 is a
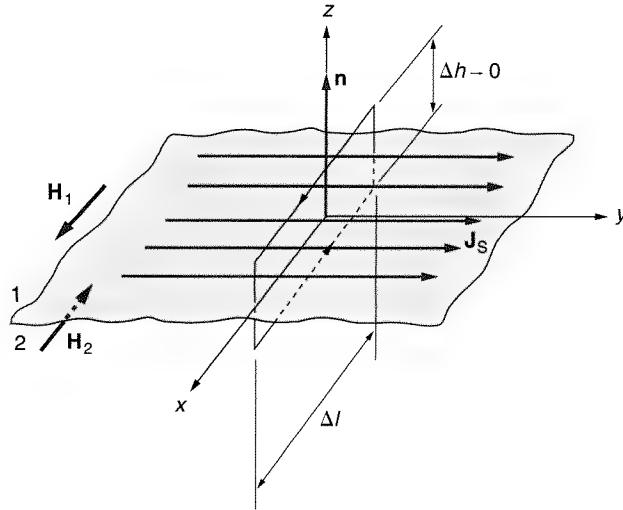


Figure 19.2 Surface current on boundary between two media. The magnetic field due to this current is locally normal to the surface-current density vector.

perfect conductor, no field is there, and we have

$$\mathbf{B}_{1norm} = \mathbf{B}_{2norm}, \quad \text{or} \quad \mathbf{B}_{norm} = 0 \quad \text{on surface of perfect conductor.} \qquad (19.13)$$

*(General boundary condition for normal components of* $\mathbf{B}$*)*

It is worthwhile repeating what we need boundary conditions for. These equations are, in fact, Maxwell's equations specialized to boundary surfaces. Therefore in a medium consisting of several bodies of different properties, the field transition from one body to the adjacent body, through a boundary surface, *must* be as required by the boundary conditions. If this were not so, such an electromagnetic field could not be a real field, because it would not satisfy the field equations *everywhere*.

*Questions and problems:*   Q19.2 to Q19.4

## 19.4   Maxwell's Equations in Differential Form

Maxwell's equations in integral form, Eqs. (19.5) to (19.8), can be transformed into a set of differential equations, known as Maxwell's equations in differential form. They can easily be obtained from the integral forms by applying the Stokes's and the divergence theorems of vector analysis. (If necessary, consult Appendix 1, Sections A1.4.6 and A1.4.7, to refresh your knowledge of these two theorems before proceeding further.)

Consider the first and second Maxwell's equations. By Stokes's theorem, the line integral of $\mathbf{E}$ in the first equation can be transformed into the flux of the vector curl $\mathbf{E} = \nabla \times \mathbf{E}$ through *any surface bounded by the contour C*. Therefore, instead of Eq. (19.5) we can write the equivalent equation

$$\int_S \nabla \times \mathbf{E} \cdot d\mathbf{S} = - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S}. \qquad (19.14)$$

The two surfaces have the same boundary contour, but they may or may not be the same. If they are the same, *any* surface bounded by $C$ can be chosen. Such an equation can be satisfied, however, only if the integrands in the two integrals are equal at all points, that is, if $\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$. Maxwell's second equation can be transformed in exactly the same manner.

The third and fourth Maxwell's equations can also be written in an equivalent form from which we can obtain their differential counterparts. For example, apply the divergence theorem to the left side of Eq. (19.7), to obtain

$$\int_v \nabla \cdot \mathbf{D} \, dv = \int_v \rho \, dv. \qquad (19.15)$$

Note that the domains $v$ on the two sides of the equation are the same. This equation can be satisfied for any domain $v$ only if the integrands are equal at all points, that is, if $\nabla \cdot \mathbf{D} = \rho$. In the same manner, we can transform the fourth Maxwell's equation.

From these derivations, Maxwell's equations in differential form read

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad (19.16)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \qquad (19.17)$$

$$\nabla \cdot \mathbf{D} = \rho \qquad (19.18)$$

$$\nabla \cdot \mathbf{B} = 0. \qquad (19.19)$$

*(Maxwell's equations in differential form)*

Let us add here the current continuity equation in differential form, obtained in the same manner as the last two equations:

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}. \qquad (19.20)$$

*(Current continuity equation in differential form)*

To these equations (as well as to their integral counterparts) it is necessary to add the relationships between vectors (1) $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{P}$; (2) $\mathbf{B}$, $\mathbf{H}$, and $\mathbf{M}$; and (3) $\mathbf{J}$ and $\mathbf{E}$:

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \qquad \mathbf{P} = \mathbf{P}(\mathbf{E}) \qquad (19.21)$$

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \qquad \mathbf{M} = \mathbf{M}(\mathbf{B}) \qquad (19.22)$$

$$\mathbf{J} = \mathbf{J}(\mathbf{E}). \qquad (19.23)$$

For linear media, which are practically the only media we consider in this text, we have

$$\mathbf{D} = \epsilon \mathbf{E} \qquad \mathbf{B} = \mu \mathbf{H} \qquad \mathbf{J} = \sigma \mathbf{E}. \qquad (19.24)$$

*(Constitutive relations for linear media)*

Equations (19.21) to (19.23) and (19.24) are often referred to as the *constitutive relations*.

The differential Maxwell's equations are used for solving many electromagnetic problems. There are modern, extremely powerful numerical methods for solving these equations directly. As computers evolve, increasingly complex electromagnetic problems can be solved numerically in a reasonable amount of time.

It is interesting that the fourth equation follows from the first. Indeed, if we take the divergence of the left-hand and right-hand sides of Eq. (19.16), the left-hand side

is equal to zero, because $\nabla \cdot (\nabla \times \mathbf{F})$ (divergence of the curl) of any vector functions $\mathbf{F}$ is identically zero. Therefore, $\partial(\nabla \cdot \mathbf{B})/\partial t = 0$, which means that $\mathbf{B}$ does not depend on time. So if at *any time in the past* $\mathbf{B} = 0$ (and therefore also $\nabla \cdot \mathbf{B}$), which certainly was the case, then $\nabla \cdot \mathbf{B} = 0$ generally. In a similar manner one can prove that with the aid of the current continuity equation, the third Maxwell's equation follows from the second.

*Questions and problems:*    Q19.5 to Q19.20

## 19.5    Maxwell's Equations in Complex (Phasor) Form

Maxwell's equations in differential form, Eqs. (19.16) to (19.19), are partial differential equations with three space coordinates and time as independent variables. Very often, the time variation of the sources is sinusoidal. *If the medium is also linear*, we know that all quantities vary in time sinusoidally. It is then possible to eliminate time from the equations, and thus simplify them. The procedure is very similar to that in circuit theory. The difference is that here we have *vector quantities* in addition to scalar quantities, and that these quantities are functions of space coordinates.

Quantities varying sinusoidally in time are often called *time-harmonic*. Their time dependence can be written in the form $\cos(\omega t + \varphi)$, where $\omega = 2\pi f$ is the angular frequency (in radians per second), $f$ is the frequency (in Hz), and $\varphi$ is the initial phase. In general, $\varphi$ is a function of coordinates. In the case of vector quantities, the initial phases of the three vector components at a point can be different.

**Example 19.3—Complex field quantities.** To understand the logic of complex representation of time-harmonic vectors, consider the $x$ component of a time-harmonic electric field of angular frequency $\omega$:

$$E_x(x, y, z, t) = E_{x\,max}(x, y, z)\, \cos[\omega t + \varphi(x, y, z)]. \tag{19.25}$$

*Euler's identity* allows us to express the cosine as a sum of complex exponentials:

$$\cos(\omega t + \varphi) = \frac{e^{j\omega t}e^{j\varphi} + e^{-j\omega t}e^{-j\varphi}}{2}, \tag{19.26}$$

where $j = \sqrt{-1}$ is the imaginary unit.

The time derivative of $E_x(x, y, z, t)$ can be written as

$$\frac{\partial}{\partial t}E_x(x, y, z, t) = E_{x\,max}(x, y, z)\, \frac{1}{2}\left(j\omega e^{j\omega t}e^{j\varphi(x,y,z)} - j\omega e^{-j\omega t}e^{-j\varphi(x,y,z)}\right). \tag{19.27}$$

All the quantities from Maxwell's equations can be expressed in this form. The equations written in such a way will contain some parts with a factor $e^{j\omega t}$, and the same parts with a factor $e^{-j\omega t}$. Because the two functions, $e^{j\omega t}$ and $e^{-j\omega t}$, are independent, the factors they multiply must be zero in order that the equations be satisfied at any $t$. In other words, instead of each equation, we get *two equivalent complex equations*. In these equations, time does not appear explicitly, and the time derivatives are replaced by $j\omega$, or $-j\omega$.

Formally, one of these complex equations can be obtained from the initial equation by replacing all the cosines with $e^{j\omega t}$, and the other by replacing the cosines with $e^{-j\omega t}$. Then, after differentiating with respect to time, all factors with $e^{j\omega t}$ and $e^{-j\omega t}$ cancel out. Although both $e^{j\omega t}$ and $e^{-j\omega t}$ can be used, it is customary in electrical engineering to replace the cosine with $e^{j\omega t}$, so that the first time derivative is replaced by the factor $j\omega$, the second time derivative by the factor $-\omega^2$, etc.

A phasor quantity in electrical engineering is written as a complex root-mean square (rms) value. To stress that a quantity is a phasor or complex, the International Electronics Commission (IEC) recommends that it be underlined, as follows:

$$\underline{A} = A(x, y, z)\, e^{j\varphi(x,y,z)} = \frac{A_{\max}(x, y, z)}{\sqrt{2}}\, e^{j\varphi(x,y,z)}. \tag{19.28}$$

The magnitude of the complex quantity is represented with the rms value instead of the maximum value because the expressions for average power and energy are conveniently expressed with rms values. Most instruments show rms values.

When dealing with complex vectors, it is important to keep in mind the following. A *real vector* has three components and, at any given moment, can be drawn as an arrow in space. The arrow describes the direction and magnitude of the vector. A *complex vector* is a set of *six* numbers, three real and three imaginary parts of its components. This is why, in general, a complex vector cannot be represented with an arrow.

After all these explanations, we can finally write down the simplest and most often quoted (but least general) form of Maxwell's equations—their complex form:

$$\nabla \times \underline{\mathbf{E}} = -j\omega\underline{\mathbf{B}}, \tag{19.29}$$

$$\nabla \times \underline{\mathbf{H}} = \underline{\mathbf{J}} + j\omega\underline{\mathbf{D}}, \tag{19.30}$$

$$\nabla \cdot \underline{\mathbf{D}} = \underline{\rho}, \tag{19.31}$$

$$\nabla \cdot \underline{\mathbf{B}} = 0. \tag{19.32}$$

*(Maxwell's equations in complex form)*

It is important to keep in mind that these equations are valid *only for linear media.* Otherwise, as explained, all quantities cannot simultaneously be time-harmonic.

In addition, we have the current continuity equation in complex form,

$$\nabla \cdot \underline{\mathbf{J}} = -j\omega\underline{\rho}, \tag{19.33}$$

*(Current continuity equation in complex form)*

as well as the constitutive relations with complex vectors (phasors), and with complex permittivity, permeability, and conductivity,

$$\underline{\mathbf{D}} = \underline{\epsilon}\,\underline{\mathbf{E}}, \qquad \underline{\mathbf{B}} = \underline{\mu}\,\underline{\mathbf{H}}, \qquad \underline{\mathbf{J}} = \underline{\sigma}\,\underline{\mathbf{E}}. \tag{19.34}$$

*(Constitutive relations in complex form)*

*Questions and problems:*  Q19.21 to Q19.24, P19.4

## 19.6  Poynting's Theorem

Poynting's theorem is the mathematical expression of the law of conservation of energy as applied to electromagnetic fields.

To obtain an energy expression from Maxwell's equations, we have to combine them in an appropriate way. We know that the expression $\mathbf{J} \cdot \mathbf{E}$ is dissipated (Joule's) power per unit volume. Note that $\mathbf{E}$ stands for the electric field due to charges and time-varying currents. In Section 10.5 we introduced the concept of the impressed electric field, $\mathbf{E}_i$. It was defined as a field equivalent to nonelectric forces acting on electric charges. Therefore the expression $\mathbf{J} \cdot \mathbf{E}_i$ is the power of impressed (external) distributed sources per unit volume.

With this in mind, consider Maxwell's differential equations (19.16) and (19.17), which we repeat for convenience:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{19.35 = 19.16}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}. \tag{19.36 = 19.17}$$

To obtain a power expression that must be satisfied by an electromagnetic field, we must combine *both* of these equations because both must simultaneously be satisfied for a real field. Let us therefore multiply (find the dot product of) the first of these equations by $\mathbf{H}$, the second by $-\mathbf{E}$, and then add the two equations thus obtained. The result is

$$\mathbf{H} \cdot \nabla \times \mathbf{E} - \mathbf{E} \cdot \nabla \times \mathbf{H} = -\mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} - \mathbf{E} \cdot \mathbf{J} - \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t}. \tag{19.37}$$

Now, from vector analysis (see Appendix 2, No. 21)

$$\mathbf{H} \cdot \nabla \times \mathbf{E} - \mathbf{E} \cdot \nabla \times \mathbf{H} = \nabla \cdot (\mathbf{E} \times \mathbf{H}). \tag{19.38}$$

If we assume the medium to be linear, we can write

$$\mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} = \mu \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} = \frac{\partial}{\partial t}\left(\frac{1}{2}\mu \mathbf{H} \cdot \mathbf{H}\right) = \frac{\partial}{\partial t}\left(\frac{1}{2}\mu H^2\right), \tag{19.39}$$

and

$$\mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} = \epsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} = \frac{\partial}{\partial t}\left(\frac{1}{2}\epsilon E^2\right). \tag{19.40}$$

For linear media, $\mathbf{J} = \sigma(\mathbf{E} + \mathbf{E}_i)$, so that $\mathbf{E} = (\mathbf{J}/\sigma - \mathbf{E}_i)$. If we substitute this expression of $\mathbf{E}$ into the term $\mathbf{E} \cdot \mathbf{J}$ in Eq. (19.37), taking into account Eqs. (19.38) to (19.40), after simple manipulations Eq. (19.37) becomes

$$\mathbf{E}_i \cdot \mathbf{J} = \frac{J^2}{\sigma} + \frac{\partial}{\partial t}\left(\frac{1}{2}\epsilon E^2 + \frac{1}{2}\mu H^2\right) + \nabla \cdot (\mathbf{E} \times \mathbf{H}). \tag{19.41}$$

Let us multiply this equation by a volume element $dv$ and integrate over an arbitrary volume $v$ of the field. The last term of the equation thus obtained is a volume integral of the divergence of the vector $(\mathbf{E} \times \mathbf{H})$. By the use of the divergence theorem, this volume integral can be transformed into a surface integral over the surface $S$ bounding the volume $v$. So we finally obtain

$$\int_v \mathbf{E}_i \cdot \mathbf{J}\, dv = \int_v \frac{J^2}{\sigma}dv + \frac{\partial}{\partial t}\int_v \left(\frac{1}{2}\epsilon E^2 + \frac{1}{2}\mu H^2\right)dv + \oint_S (\mathbf{E} \times \mathbf{H}) \cdot d\mathbf{S}. \tag{19.42}$$
$$\textit{(Poynting's theorem)}$$

This is *Poynting's theorem*. It tells us about power balance inside a volume $v$ of the electromagnetic field.

The term on the left represents the power of all the sources inside $v$. The terms on the right show how this power is used. One part (represented by the first term) is transformed inside $v$ into heat. The other part (represented by the second term) is used to change (increase if positive, decrease if negative) the energy localized in the electric and magnetic field inside $v$. Because we consider a finite volume of the field, we need a term representing possible exchange of energy with the rest of the field, through the boundary of $v$, that is, surface $S$. According to Poynting, the last term on the right has precisely that meaning:

$$\oint_S (\mathbf{E} \times \mathbf{H}) \cdot d\mathbf{S} = \text{power transferred through } S \text{ to a region outside } S. \tag{19.43}$$

This statement is also frequently considered as Poynting's theorem.

According to Poynting's theorem, the cross product $(\mathbf{E} \times \mathbf{H})$ can be interpreted as the power transferred by the electromagnetic field per unit area. The direction of the vector $(\mathbf{E} \times \mathbf{H})$ then shows the direction of transfer of energy through a surface perpendicular to it. The vector $(\mathbf{E} \times \mathbf{H})$ is referred to as the *Poynting vector*. We will designate it by $\mathcal{P}$ (calligraphic $P$):

$$\mathcal{P} = \mathbf{E} \times \mathbf{H} \qquad (\text{W/m}^2). \tag{19.44}$$

*(Definition of the Poynting vector)*

The unit of Poynting's vector is $\text{W/m}^2$ (watts per square meter).

Poynting's theorem, as a mathematical expression of the law of conservation of energy in the electromagnetic field, is an extremely useful theorem. Note, however, that it is valid only for electromagnetic fields that are described *simultaneously* by the first and second of Maxwell's equations. The following example shows that in other cases Poynting's theorem does not make sense.

**Example 19.4—Formal application of Poynting's theorem to crossed electrostatic and magnetostatic fields.** Consider the system shown in Fig. 19.3. A charged parallel-plate capacitor and a permanent magnet are positioned so that their fields (electrostatic and magnetostatic) overlap. Consequently, considered formally, Poynting's vector in the figure is directed into the page. This could be interpreted as if energy is perpetually circulating through this region, and the only problem is how to capture it. This reasoning, however, is not correct. These electric and magnetic fields *are not coupled* (we can move the magnet, for example, without affecting the electric field of the capacitor). Combining the two fields in this case is like combining potatoes and oranges.

**Example 19.5—Energy transfer through a coaxial cable.** The cross section of a coaxial cable is sketched in Fig. 19.4. Assume that the voltage between the cable conductors is $V$, and that there is a dc current in the cable of intensity $I$, as indicated. It is a simple matter to conclude that the generator is connected in the direction toward the reader, and the load away from the reader. The Poynting vector is directed away from the reader. According to the interpretation of the Poynting vector, this means that energy is flowing through the cable away from the generator, as it should.
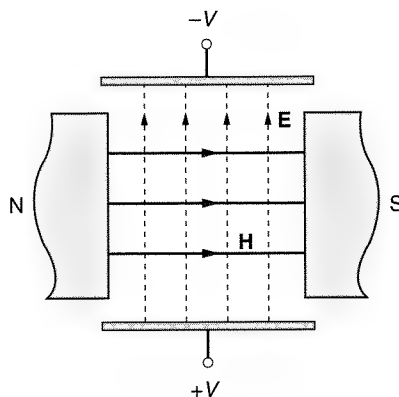


Figure 19.3 Crossed electrostatic and magnetostatic fields
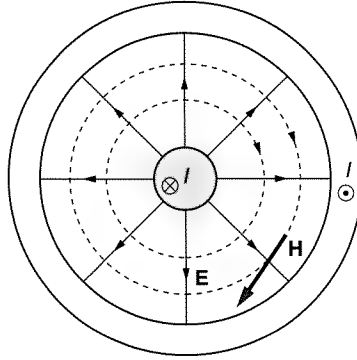
Figure 19.4 Cross section of a coaxial cable
with lines of vectors **E** and **H**

It is left as an exercise for the reader to prove that the flux of the Poynting vector through the cross section of the cable equals exactly $VI$. Note that the intensity of the Poynting vector is the largest at the inner conductor surface, which means that most of the power flows near that surface.

**Example 19.6—Poynting's theorem in complex form.** Starting from the complex form of Maxwell's equations, it is not difficult to obtain Poynting's theorem in complex form. The principal difference of the derivation is that we start from the complex form of the first equation, from the *complex conjugate form* of the second equation, and multiply the first equation (find the dot product) with the *complex conjugate*, $\underline{\mathbf{H}}^*$, of the vector $\underline{\mathbf{H}}$. The rest of the derivation is quite similar to that given for Poynting's theorem for arbitrary time dependence, and it is left as an exercise for the reader. The result is

$$\int_v \underline{\mathbf{E}}_i \cdot \underline{\mathbf{J}}^* dv = \int_v \frac{J^2}{\sigma} dv + 2j\omega \int_v \left( \frac{1}{2}\mu H^2 - \frac{1}{2}\epsilon E^2 \right) dv + \oint_S (\underline{\mathbf{E}} \times \underline{\mathbf{H}}^*) \cdot d\mathbf{S}. \qquad (19.45)$$

*(Poynting's theorem in complex form)*

This is Poynting's theorem in complex form. The vector

$$\underline{\mathcal{P}} = \underline{\mathbf{E}} \times \underline{\mathbf{H}}^* \qquad (\text{W/m}^2) \qquad\qquad\qquad\qquad (19.46)$$

*(The complex Poynting vector)*

is known as the *complex Poynting vector*.

The equation expressing the Poynting theorem in complex form has a real and an imaginary part. It is left to the reader as an exercise to write these two parts of the equation and to discuss their meaning.

*Questions and problems:*   Q19.25 to Q19.30, P19.5 to P19.9

## 19.7    The Generalized Definition of Conductors and Insulators

For linear media and time-harmonic variation of the fields, it is possible to clearly distinguish what a good conductor and a good insulator are. Let a time-harmonic electromagnetic field of angular frequency $\omega$ exist in a medium of permittivity $\epsilon$ and conductivity $\sigma$. The second Maxwell's equation in complex form becomes

$$\nabla \times \mathbf{H} = (\sigma + j\omega\epsilon)\mathbf{E}. \tag{19.47}$$

For a perfect dielectric, $\sigma$ in this equation does not exist. For a very good conductor, displacement current is negligible, so the term $j\omega\epsilon$ is missing. Thus, at a frequency $f = \omega/(2\pi)$, we can define a good conductor by the inequality

$$\sigma \gg \omega\epsilon \quad \text{(definition of a good conductor),} \tag{19.48}$$

and a good insulator by the inequality

$$\sigma \ll \omega\epsilon \quad \text{(definition of a good insulator).} \tag{19.49}$$

## 19.8    The Lorentz Potentials

In Chapter 4 we introduced the concept of the electric scalar potential. This is just one in a family of potentials used in the analysis of electromagnetic fields. A potential is an auxiliary scalar or vector function, which is usually easier to calculate than the field vectors themselves, and from which the field vectors are obtained in some simple manner, usually by differentiation.

We will introduce here a pair of potentials that seem to be used most often in electromagnetic field analysis. One of these is the generalized scalar potential we already know. The other is a vector function, known as the *magnetic vector potential*. The specific pair of potentials we will now derive are known as the *Lorentz potentials*. For reasons to become apparent later, they are also known as the *retarded potentials*.

Note first that $\nabla \cdot (\nabla \times \mathbf{A}) = 0$ for any vector function $\mathbf{A}$ (see Appendix 2, No. 24). Since $\nabla \cdot \mathbf{B} = 0$, it follows that it is always possible to express the magnetic flux density vector $\mathbf{B}$ as

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{19.50}$$

*(Definition of magnetic vector potential)*

The vector function $\mathbf{A}$ is known as the *magnetic vector potential*.

If the expression for $\mathbf{B}$ in Eq. (19.50) is introduced into the first Maxwell's equation, we obtain

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{A}). \tag{19.51}$$

$\int^{A1.47}$

This means that $\nabla \times (\mathbf{E} + \partial \mathbf{A}/\partial t) = 0$. Now, we know that $\nabla \times (\nabla V) = 0$ always [see Appendix 1, Eq. (A1.50)]. Therefore Eq. (19.51) implies that $(\mathbf{E} + \partial \mathbf{A}/\partial t) = -\nabla V$, *and not zero*. (The negative gradient is used for convenience.) Thus the electric field strength can be expressed as

$$\mathbf{E} = -\nabla V - \frac{\partial \mathbf{A}}{\partial t}. \tag{19.52}$$

*(Electric field strength in terms of retarded potentials)*

Evidently, for time-invariant fields $V$ becomes the electric scalar potential we know. Therefore we retain the same name for $V$ in this case, where $V$ is an arbitrary function of time.

So we have two equations, (19.50) and (19.52), from which we can easily calculate vectors $\mathbf{E}$ and $\mathbf{B}$, provided we know the two potentials, $V$ and $\mathbf{A}$. For obtaining Eqs. (19.50) and (19.52) we used the first and the fourth Maxwell's equations. For determining these potentials in terms of the field sources, $\rho$ and $\mathbf{J}$, we therefore make use of the other two Maxwell's equations.

Let us assume that the medium is linear and homogeneous, of permittivity $\epsilon$ and permeability $\mu$. Then, substituting Eqs. (19.50) and (19.52) into the second and third Maxwell's equation, we obtain, respectively,

$$\nabla \times (\nabla \times \mathbf{A}) = \mu \mathbf{J} - \epsilon \mu \frac{\partial}{\partial t}(\nabla V) - \epsilon \mu \frac{\partial^2 \mathbf{A}}{\partial t^2}, \tag{19.53}$$

and

$$\nabla \cdot (\nabla V) = \nabla^2 V = -\frac{\rho}{\epsilon} - \nabla \cdot \frac{\partial \mathbf{A}}{\partial t}. \tag{19.54}$$

Since $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$ [see Appendix 1, Eq. (A1.37)], Eq. (19.53) becomes

$$\nabla^2 \mathbf{A} = -\mu \mathbf{J} + \epsilon \mu \, \nabla \frac{\partial V}{\partial t} + \epsilon \mu \frac{\partial^2 \mathbf{A}}{\partial t^2} + \nabla(\nabla \cdot \mathbf{A}). \tag{19.55}$$

There is a theorem in vector analysis called the *Helmholtz theorem*. It says that a vector function is uniquely defined if its curl and divergence are known at every point in space. We already know what the curl of $\mathbf{A}$ is ($\nabla \times \mathbf{A} = \mathbf{B}$), so we need to define its divergence in order that it be unique. Because only $\nabla \times \mathbf{A}$ matters ($\mathbf{B} = \nabla \times \mathbf{A}$), we can define $\nabla \cdot \mathbf{A}$ in an infinite number of ways, resulting in an infinite number of pairs of potentials $\mathbf{A}$ and $V$. Having this freedom of choice, it is wise to adopt $\nabla \cdot \mathbf{A}$ so that we can solve Eqs. (19.54) and (19.55) most easily.

It is a simple matter to conclude that if we adopt the *Lorentz condition* for $\nabla \cdot \mathbf{A}$,

$$\nabla \cdot \mathbf{A} = -\epsilon \mu \frac{\partial V}{\partial t}, \tag{19.56}$$

*(The Lorentz condition)*

Eqs. (19.54) and (19.55) take the simplest possible forms, each becoming a partial differential equation in a single unknown, $V$ in the first case and $\mathbf{A}$ in the second case:

$$\nabla^2 V - \epsilon\mu\frac{\partial^2 V}{\partial t^2} = -\frac{\rho}{\epsilon}, \tag{19.57}$$

$$\nabla^2 \mathbf{A} - \epsilon\mu\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu\mathbf{J}. \tag{19.58}$$

Because the $x$ component of the last equation in a rectangular coordinate system is given by [see Appendix 1, Eq. (A1.39)]

$$\nabla^2 A_x - \epsilon\mu\frac{\partial^2 A_x}{\partial t^2} = -\mu J_x, \tag{19.59}$$

and similarly for the $y$ and $z$ components, we need to solve only Eq. (19.57). The solution of Eq. (19.58) will then be obtained as a vector sum of analogous solutions for the vector components of $\mathbf{A}$.

Solving Eq. (19.57) is not simple and does not add anything to the understanding of the final result. We therefore give only the final result:

$$V(\mathbf{r}, t) = \frac{1}{4\pi\epsilon}\int_{v'}\frac{\rho(\mathbf{r}', t - R/c)}{R}\mathrm{d}v' \qquad c = \frac{1}{\sqrt{\epsilon\mu}}. \tag{19.60}$$

So the solution of Eq. (19.58) is

$$\mathbf{A}(\mathbf{r}, t) = \frac{\mu}{4\pi}\int_{v'}\frac{\mathbf{J}(\mathbf{r}', t - R/c)}{R}\mathrm{d}v' \qquad c = \frac{1}{\sqrt{\epsilon\mu}}. \tag{19.61}$$

These are the *Lorentz potentials*. The meaning of $\mathbf{r}$, $\mathbf{r}'$, and $\mathbf{R}$ is illustrated in Fig. 19.5.
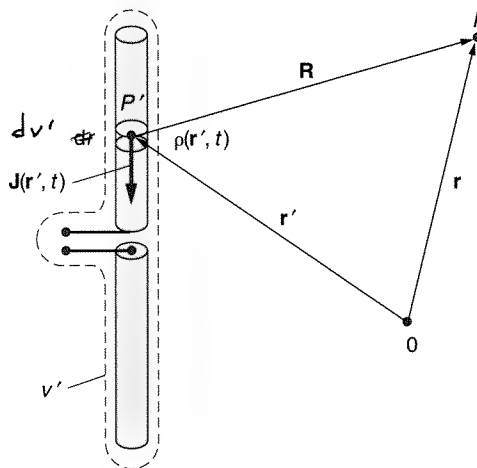


Figure 19.5 Explanation of symbols in Eqs. (19.60) and (19.61)

We stress again that the entire derivation, and therefore also Eqs. (19.60) and (19.61), is valid only for *homogeneous and linear media*.

What is the physical meaning of the expressions for the potentials in Eqs. (19.60) and (19.61)? Say there is an elemental source at a point $P'$ whose position vector is $\mathbf{r}'$, as in Fig. 19.5. We are observing the fields due to this source at a point $P$ defined by the position vector $\mathbf{r}$. The magnitude of the field at point $P$ at a time $t$ is not the one that the source produces at time $t$, but at an earlier time, $t - R/c$. In other words, in a homogeneous dielectric of permittivity $\epsilon$ and permeability $\mu$, the fields propagate with a finite velocity, $c = 1/\sqrt{\epsilon\mu}$, that is, they are *retarded* in reaching the field point. For this reason, the Lorentz potentials are often termed the *retarded potentials*.

In the case of a vacuum ($\epsilon = \epsilon_0$, $\mu = \mu_0$), the velocity $c$ of propagation of the potentials *becomes exactly the speed of light in a vacuum*, $c_0$, a calculation left as an exercise for the reader.

**Example 19.7—Retarded potentials in complex (phasor) form.** Very often, sources of an electromagnetic field are time-harmonic. In that case the retarded potentials can be written without explicit time dependence. The procedure for obtaining the complex potentials is simple—we just assume the sources, $\rho$ and $\mathbf{J}$, and the potentials to vary following the law $e^{j\omega t}$. So we obtain

$$\underline{V}(\mathbf{r}) = \frac{1}{4\pi\epsilon} \int_{v'} \frac{\rho(\mathbf{r}')e^{-j\omega R/c}}{R}\,dv', \qquad c = \frac{1}{\sqrt{\epsilon\mu}}, \tag{19.62}$$

*(Complex retarded scalar potential)*

and

$$\underline{\mathbf{A}}(\mathbf{r}) = \frac{\mu}{4\pi} \int_{v'} \frac{\mathbf{J}(\mathbf{r}')e^{-j\omega R/c}}{R}\,dv' \qquad c = \frac{1}{\sqrt{\epsilon\mu}}. \tag{19.63}$$

*(Complex retarded vector potential)*

**Example 19.8—Definition of quasi-static fields.** It is interesting that for time-harmonic fields it is possible to inspect whether a field in a system can be considered practically as a static field (or a *quasi-static field*), or not.

From Eqs. (19.62) and (19.63) we see that the retardation can be neglected provided that the largest dimension of the field domain we consider, $d_{max}$, is determined by the inequality $\omega d_{max}/c \ll 1$, or

$$d_{max} \ll \frac{c}{\omega} = \frac{1}{\omega\sqrt{\epsilon\mu}} \qquad \text{(the condition for quasi-static fields)}.$$

*Questions and problems:*    Q19.31 to Q19.36, P19.10 to P19.13

# 19.9  Chapter Summary

1. The general equations of the electromagnetic field, known as Maxwell's equations, are as fundamental in electromagnetic field theory as are Newton's laws in mechanics.

**2.** Some general consequences of Maxwell's equations include general boundary conditions, the Poynting theorem (the theorem on the transfer of energy by the electromagnetic field), and the possibility of making a clear distinction between conductors and insulators for time-harmonic fields.

**3.** Field vectors can be expressed in terms of auxiliary functions, called potentials.

**4.** The expressions for potentials indicate that the speed of electromagnetic disturbances in a vacuum is that of light.

**5.** For sinusoidal field variation, the expressions for the potentials in complex form enables us to define the dimensions of systems in which fields can be considered approximately as static (quasi-static fields).

# QUESTIONS

**Q19.1.** Why (and when) is it allowed to move the time derivative in Eq. (19.1) to act on **D** only, and thus obtain Eq. (19.2)?

**Q19.2.** Does Eq. (19.5) tell us that a time-varying magnetic field is the source of a time-varying electric field? Explain.

**Q19.3.** Why would an electric field inside a perfect conductor produce a current of infinite density? Would such a current be physically possible? Explain.

**Q19.4.** Why are surface currents possible on surfaces of perfect conductors, when a nonzero tangential electric field there is not possible? Is this a current of finite volume density?

**Q19.5.** Write the full set of Maxwell's equations in differential form for the special case of a static electric field, assuming that the dielectric is linear, but inhomogeneous.

**Q19.6.** Write the full set of Maxwell's equations in differential form for the special case of a static electric field produced by the charges on a set of conducting bodies situated in a vacuum.

**Q19.7.** Write the full set of Maxwell's equations in differential form for the special case of a steady current flow in a homogeneous conductor of conductivity $\sigma$, with no impressed electric field.

**Q19.8.** Write the full set of Maxwell's equations in differential form for the special case of a steady current flow in an inhomogeneous poor dielectric, with impressed electric field $\mathbf{E}_i$ present.

**Q19.9.** Write the full set of Maxwell's equations in differential form for the special case of a time-constant magnetic field in a linear medium of permeability $\mu$, produced by a steady current flow.

**Q19.10.** Write the full set of Maxwell's equations in differential form for the special case of a time-constant magnetic field, produced by a permanent magnet of magnetization **M** (a function of position).

**Q19.11.** Write the full set of Maxwell's equations in differential form for the special case of a time-constant magnetic field produced by both steady currents and magnetized matter, if the medium is not linear.

**Q19.12.** Write the full set of Maxwell's equations in differential form for the special case of a quasi-static electromagnetic field, produced by quasi-static currents in nonferromagnetic conductors.

**Q19.13.** Write Maxwell's equations in differential form for an arbitrary electromagnetic field in a vacuum, no free charges being present.

**Q19.14.** Write Maxwell's equations for an arbitrary electromagnetic field in a homogeneous perfect dielectric of permittivity $\epsilon$ and permeability $\mu$.

**Q19.15.** Write the full set of differential Maxwell's equations *in scalar form* in the rectangular coordinate system. Note that *eight* simultaneous, partial differential equations result. Write these equations neatly and save them for future reference.

**Q19.16.** Repeat question Q19.15 for the cylindrical coordinate system.

**Q19.17.** Repeat question Q19.15 for the spherical coordinate system.

**Q19.18.** Write differential Maxwell's equations in scalar form for the particular case of an electromagnetic field in a vacuum ($\mathbf{J} = 0$, $\rho = 0$), if the field vectors are only functions of the cartesian coordinate $z$ and of time $t$.

**Q19.19.** Write differential Maxwell's equations in scalar form for a good conductor, for the particular case of an axially symmetrical system with dependence of the field vectors only on the cylindrical coordinate $r$ and time $t$. Assume that $\mathbf{J} = J_z \mathbf{u}_z$ and $\rho = 0$.

**Q19.20.** Repeat question Q19.19 for $\mathbf{B} = B_z \mathbf{u}_z$.

**Q19.21.** Write differential Maxwell's equations in complex form for an arbitrary electromagnetic field in a very good conductor, of conductivity $\sigma$ and permeability $\mu$.

**Q19.22.** Write differential Maxwell's equations in complex form for a quasi-static electromagnetic field.

**Q19.23.** Write differential Maxwell's equations in complex form for an arbitrary electromagnetic field in a perfect dielectric of permittivity $\epsilon$ and permeability $\mu$, no free charges being present.

**Q19.24.** Write the most general integral Maxwell's equations in complex form.

**Q19.25.** The current intensity through a resistor of resistance $R$ is $I$. What is the flux of the Poynting vector through any closed surface enclosing the resistor?

**Q19.26.** A capacitor, of capacitance $C$, is charged with a charge $Q$. What is the flux of the Poynting vector through any surface enclosing the capacitor, if the charge $Q$ (1) is constant in time, or (2) varies in time as $Q = Q_m \cos \omega t$?

**Q19.27.** A coil, of inductance $L$, carries a current $i(t)$. What is the flux of the Poynting vector through any surface enclosing the coil?

**Q19.28.** A dc generator of emf $\mathcal{E}$ is open-circuited. What is the flux of the Poynting vector through any surface enclosing the generator?

**Q19.29.** Repeat question Q19.28 assuming that a current $i(t)$ flows through the generator, and its internal resistance is $R$.

**Q19.30.** What is the time-average value of the Poynting vector, if complex rms values are known for the electric and magnetic field strength, $\mathbf{E}$ and $\mathbf{H}$?

**Q19.31.** The largest dimension of a coil at a very high frequency is on the order of $(\omega \sqrt{\epsilon \mu})^{-1}$. Is it possible at such high frequencies to define the inductance of the coil in the same way as in a quasi-static case? Explain.

**Q19.32.** The length of a long 60-Hz power transmission line is equal to $0.5(\omega \sqrt{\epsilon_0 \mu_0})^{-1}$. Is this a quasi-static system? What is the length of the line?

**Q19.33.** A parallel-plate capacitor has plates of linear dimensions comparable with $(\omega \sqrt{\epsilon \mu})^{-1}$, where $\omega$ is the operating angular frequency. Is it possible to determine the capacitance of such a capacitor in the same way as in the static and quasi-static case? Explain.

**Q19.34.** An electric circuit operates at a high frequency $f$. The largest linear dimension of the circuit is $2(\omega\sqrt{\epsilon\mu})^{-1}$. Are Kirchhoff's laws applicable in this case for analyzing the circuit? Explain.

**Q19.35.** Write the Lorentz condition in complex form.

**Q19.36.** A current pulse of duration $\Delta t = 10^{-9}$ s was excited in a small wire loop. After how many $\Delta t$'s is the magnetic and induced electric field of this pulse going to be detected at a point $r = 10$ m from the loop?

# PROBLEMS

**P19.1.** A current $i(t) = I_m \cos \omega t$ flows through the leads of a parallel-plate capacitor of plate area $S$ and distance between them $d$. If the permittivity of the dielectric of the capacitor is $\epsilon$, prove that the displacement current through the capacitor dielectric is exactly $i(t)$. Ignore fringing effects.

**P19.2.** A spherically symmetrical charge distribution disperses under the influence of mutually repulsive forces. Suppose that the charge density $\rho(r, t)$, as a function of the distance $r$ from the center of symmetry and of time, is known. Prove that the total current density at any point is zero.

**P19.3.** Determine the magnetic field as a function of time for the dispersing charge distribution in problem P19.2.

**\*P19.4.** Small-scale models are used often in engineering practice, including electrical engineering. Starting from differential Maxwell's equations for a linear medium, derive the necessary conditions for the electromagnetic field in a small-scale model to be similar to the field in a real, $n$ times larger model. (These conditions are usually referred to as the conditions of the *electrodynamic similitude*.) (*Hints:* (1) Write the first two differential Maxwell's equations for the full-scale system, and for the model. (2) Note that the coordinates in the latter are $n$ times smaller, and find the conditions under which, in spite of that, the two sets of equations will be the same.)

**P19.5.** A lossless coaxial cable, of conductor radii $a$ and $b$, carries a steady current of intensity $I$. The potential difference between the cable conductors is $V$. Prove that the flux of the Poynting vector through a cross section of the cable is $VI$, using the known expressions for vectors **E** and **H** in the cable. Sketch the dependence of the magnitude of the Poynting vector on the distance $r$ from the cable axis, where $a < r < b$.

**P19.6.** Repeat problem P19.5 for an air stripline with strips of width $a$ that are a distance $d$ apart, if the current in the strips is $I$ and voltage between them $V$. Neglect the edge effects.

**P19.7.** The stripline from the preceding problem is connected to a sinusoidal generator of emf $\mathcal{E}$ and angular frequency $\omega$. The other end of the line is connected to a capacitor of capacitance $C$. Apply Poynting's theorem in complex form to a closed surface enclosing (1) the generator, or (2) the capacitor.

**P19.8.** Repeat problem P19.7 assuming that the load is an inductor of inductance $L$, instead of a capacitor.

**P19.9.** Repeat problem P19.7, assuming that the line is a lossless coaxial line of conductor radii $a$ and $b$.

**P19.10.** Derive Eqs. (19.57) and (19.58) from Eqs. (19.54), (19.55), and (19.56).

**P19.11.** Derive the retarded potentials in Eqs. (19.62) and (19.63) from Eqs. (19.60) and (19.61).

**P19.12.** Suppose a system is regarded as approximately quasi-static if its largest dimension $d$ satisfies the inequality $d\omega\sqrt{\epsilon\mu} \leq 0.1$. Determine the largest value of $d$ thus defined for the electrodynamic systems in a vacuum if the frequency of the generators is (1) 60 Hz, (2) 10 MHz, or (3) 10 GHz.

**P19.13.** Compare the rms values of vectors $\mathbf{J}$ and $\partial\mathbf{D}/\partial t$ in copper, seawater, and wet ground, for frequencies $f$ of (1) 60 Hz, (2) 10 kHz, (3) 100 MHz, or (4) 10 GHz. For copper, assume $\epsilon = \epsilon_0$, $\sigma = 56 \cdot 10^6$ S/m. For seawater, adopt $\epsilon = 10\epsilon_0$, $\sigma = 4$ S/m, and for the ground $\epsilon = 10\epsilon_0$ and $\sigma = 10^{-2}$ S/m.

# 20

# *The Skin Effect*

## 20.1 Introduction

We are now in a position to formulate any electromagnetic problem in terms of Maxwell's equations. This chapter deals with the skin effect, the first practical electromagnetic problem we will solve as an example of this kind. A related effect, the proximity effect, is then considered briefly.

We know that a time-invariant current in a homogeneous cylindrical conductor is distributed uniformly over the conductor cross section. If the conductor is not cylindrical, the time-invariant current in it is not distributed uniformly, but it exists in the *entire* conductor. We shall see in this chapter that a time-varying current has a tendency to concentrate near the surfaces of conductors. If the frequency is very high, the current is restricted to a very thin layer near the conductor surfaces, practically on the surfaces themselves. Because of this extreme case, the entire phenomenon of nonuniform distribution of time-varying currents in conductors is known as the *skin effect*.

The cause of the skin effect is electromagnetic induction. A time-varying magnetic field is accompanied by a time-varying induced electric field, which in turn creates secondary time-varying currents (induced currents) and a secondary magnetic field. We know from Lentz's law that the induced currents produce the magnetic flux, which is opposite to the external flux (which "produced" the induced currents), so that the total flux is reduced. The larger the conductivity, the larger the induced currents are, and the larger the permeability, the more pronounced is the flux reduction.

Consequently, both the total time-varying magnetic field and induced currents inside conductors are reduced when compared with the dc case.

The skin effect is of considerable practical importance. For example, at very high frequencies a very thin layer of conductor carries most of the current, so we can coat any conductor with silver (the best available conductor) and have practically the entire current flow through this thin silver coating. (Unfortunately silver oxidizes easily, so gold is often used instead because it is inert.) Even at low, power-line frequencies (60 Hz in the United States and Canada, and 50 Hz in Europe), in the case of high currents the use of thick, solid conductors is not efficient; bundled conductors are used instead.

The skin effect exists in all conductors, but as mentioned, the tendency of current and magnetic flux to be restricted to a thin layer on the conductor surface is much more pronounced for a ferromagnetic conductor than for a nonferromagnetic conductor of the same conductivity. For example, for iron at 60 Hz the thickness of this layer is on the order of only 0.5 mm. Consequently, solid ferromagnetic cores for alternating current electric motors, generators, transformers, etc., would have very high eddy-current losses. Therefore laminated cores made of thin, mutually insulated sheets are used instead. At very high frequencies, ferrites (ferrimagnetic materials) are used because they have very low conductivity when compared to metallic ferromagnetic materials.

*Questions and problems:*   Q20.1 to Q20.10

## 20.2   Skin Effect

Consider an idealized case of a sinusoidal current in a homogeneous conducting half-space, as sketched in Fig. 20.1. Let the angular frequency of the current be $\omega$ and let the medium have a conductivity $\sigma$ and permeability $\mu$. Finally, assume that the
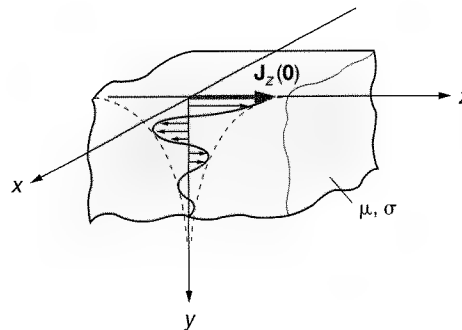


**Figure 20.1** Homogeneous conducting half-space with sinusoidal current. The amplitude of the current density vector at an instant of time versus the distance $y$ from the boundary surface is as indicated.

current density vector is parallel to the boundary surface, and that it has a single component, for example, $\mathbf{J} = J_z\mathbf{u}_z$, depending on the coordinate $y$ (the distance from the interface) only. We wish to determine the distribution of current in the conducting half-space.

At first glance, one might be tempted to think this problem is purely academic. We will see, however, that it has important practical implications. After solving Maxwell's equations, we will find that the intensity of the current density vector and of all the field vectors decreases exponentially with the distance from the boundary surface. This decrease is more rapid at higher frequencies and for higher conductivities and permeabilities. For conductors used in everyday practice (copper, for instance), and frequencies higher than about 1 MHz, the thickness of the current layer becomes less than a fraction of a millimeter. If we consider *any* conductor whose radius of curvature is much larger than the current layer thickness, the results we will obtain can be applied with high accuracy. Therefore this section has considerable practical importance and applicability.

We start the analysis from the differential form of Maxwell's equations in complex form. Because we assume the medium to be a good conductor, the displacement current density in the second equation can be neglected. We start from

$$\nabla \times \mathbf{E} = -\mathrm{j}\omega\mathbf{B} \qquad \nabla \times \mathbf{H} = \mathbf{J}. \tag{20.1}$$

For simplicity we do not underline the complex vectors $\mathbf{E}$, $\mathbf{B}$, $\mathbf{H}$, and $\mathbf{J}$. Since $\mathbf{E} = \mathbf{J}/\sigma$ and $\mathbf{H} = \mathbf{B}/\mu$, Eqs. (20.1) become

$$\nabla \times \mathbf{J} = -\mathrm{j}\omega\sigma\,\mathbf{B} \qquad \nabla \times \mathbf{B} = \mu\mathbf{J}. \tag{20.2}$$

We assumed that the current density vector has only a $z$ component, which depends only on $y$. From the Biot-Savart law and symmetry it therefore follows that there is only an $x$ component of the vector $\mathbf{B}$. According to the expression for the curl in a rectangular coordinate system, Eqs. (20.2) become

$$\frac{\mathrm{d}J_z}{\mathrm{d}y} = -\mathrm{j}\omega\sigma B_x \qquad -\frac{\mathrm{d}B_x}{\mathrm{d}y} = \mu J_z. \tag{20.3}$$

We use ordinary derivatives (not partial derivatives) because $J_z$ and $B_x$ depend only on $y$.

From Eqs. (20.3) we can eliminate $B_x$ to obtain an equation in $J_z$:

$$\frac{\mathrm{d}^2 J_z}{\mathrm{d}y^2} = \mathrm{j}\omega\mu\sigma J_z. \tag{20.4}$$

This equation has a simple solution,

$$J_z(y) = J_1\mathrm{e}^{Ky} + J_2\mathrm{e}^{-Ky}, \tag{20.5}$$

where

$$K = \sqrt{\mathrm{j}\omega\mu\sigma} = (1+\mathrm{j})\sqrt{\frac{\omega\mu\sigma}{2}} = (1+\mathrm{j})k \qquad k = \sqrt{\frac{\omega\mu\sigma}{2}}. \tag{20.6}$$

Assume that for $y = 0$ the current density is $J_z(0)$. For $y \to \infty$, the current density cannot increase indefinitely, so $J_1 = 0$. Thus we finally have

$$J_z(y) = J_z(0)e^{-ky}e^{-jky}. \tag{20.7}$$

The intensity of the current density vector decreases exponentially with increasing $y$. At a distance

$$\delta = \frac{1}{k} = \sqrt{\frac{2}{\omega\mu\sigma}} \quad \text{(m)}, \tag{20.8}$$

*(Definition of skin depth)*

the amplitude of the current density vector decreases to $1/e$ of its value $J_z(0)$ at the boundary surface. This distance is known as the *skin depth*.

As mentioned, although derived for a special case of currents in a half-space, the preceding analysis is valid for a current distribution in any conductor whose radius of curvature is much larger than the skin depth.

**Example 20.1—Skin depth for some common materials.** As an illustration, let us determine the skin depth for copper ($\sigma = 57 \cdot 10^6$ S/m, $\mu = \mu_0$), iron ($\sigma = 10^7$ S/m, $\mu_r = 1000$), seawater ($\sigma = 4$ S/m, $\mu = \mu_0$), and wet soil ($\sigma = 0.01$ S/m, $\mu = \mu_0$) at 60 Hz (power-line frequency), $10^3$ Hz, $10^6$ Hz, and $10^9$ Hz. The results are summarized in Table 20.1. Note that for iron the skin depth is very small (significantly less than a millimeter) even at the low power-line frequency. For seawater, the power-frequency skin depth is also relatively small (about 35 m), and for a radio frequency of 1 MHz it is less than 25 cm. For copper, at 1 MHz the skin depth is less than one-tenth of a millimeter.

**Example 20.2—Why not use cheap iron instead of expensive copper for distributing electric power?** The skin depth for iron at 60 Hz in Table 20.1 answers an important question. If iron has a conductivity that is only about six times less than that of copper, and copper is much more expensive than iron, why do we not use iron wires to carry electric power to our homes? With the millions of kilometers of such wires, that would mean very large savings.

Unfortunately, due to its large relative permeability, iron has a very small skin depth at powerline frequency, so the losses in iron wire are large, outweighing the savings. Thus we have to use copper or aluminum.

TABLE 20.1 Skin depth ($\delta$) for some common materials

| Material | $f = 60$ Hz | $f = 10^3$ Hz | $f = 10^6$ Hz | $f = 10^9$ Hz |
|---|---|---|---|---|
| Copper | 8.61 mm | 2.1 mm | 0.067 mm | 2.11 $\mu$m |
| Iron | 0.65 mm | 0.16 mm | 5.03 $\mu$m | 0.016 $\mu$m |
| Seawater | 32.5 m | 7.96 m | 0.25 m | 7.96 mm |
| Wet soil | 650 m | 159 m | 5.03 m | 0.16 m |

**Example 20.3—Mutual inductance between cables laid on the bottom of the sea.** Assume we have three single-phase 60-Hz cables laid at the bottom of the sea (for example, to supply electric power to an island). The cables are spaced by a few hundred meters and are parallel. (Three distant single-phase instead of one three-phase cable are often used for safety reasons: if a ship accidentally pulls and breaks one cable with an anchor, two are left. In addition, usually a spare single-phase cable is laid to enable quick replacement of a damaged one.) If the length of the cables is long (in practice, it can be many kilometers), we might expect very large mutual inductance between these cables, due to the huge loops they form. The skin depth of seawater at 60 Hz (Table 20.1), however, tells us that there will be practically *no* mutual inductance between the cables.

According to Eq. (20.7), with increasing $y$ the current density changes not only in amplitude *but also in phase*. Thus, at a distance $y = \pi/k$ from the boundary surface the vector **J** has at all times *the opposite actual direction* to that near the boundary surface. The distribution of current density as a function of $y$ at an instant is sketched in Fig. 20.1.

An important problem that we are now ready to solve is Joule's losses in the conductor per unit area of the boundary surface. Because we know the current density vector, one possibility is to integrate $[|J_z^2(y)|/\sigma]\,dy$ from $y = 0$ to infinity, which is not too difficult to do. There is an easier way, however, that does not require integration but uses the concept of the Poynting vector. This derivation is given as problem P20.7 at the end of the chapter. Here we quote and discuss the final result of this derivation. It is found that the power of Joule's losses and the internal reactive power inside the conductor, per area $S$, are given by

$$P_J = \int_S R_s |H_0|^2 \, dS = (P_{\text{reactive}})_{\text{internal}} \quad (W), \tag{20.9}$$
*(Evaluation of Joule's losses and reactive power in conductors at high frequencies)*

where $H_0$ is the complex rms value of the tangential component of the vector **H** *on the conductor surface*. (By assumption, the normal component of **H** does not exist.) $R_s$ is the *surface resistance* of the conductor, given by

$$R_s = \sqrt{\frac{\omega\mu}{2\sigma}} \quad (\Omega). \tag{20.10}$$
*(Definition of surface resistance)*

This formula for the surface resistance can be obtained if we consider the following rough approximation, illustrated on the square metal slab in Fig. 20.2. We assume that the entire high-frequency current is flowing uniformly over the cross
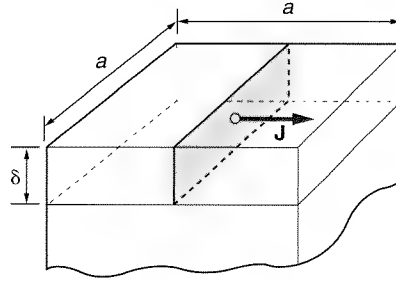
**Figure 20.2** The surface resistance of a conductive slab with a uniform current flowing through a cross section $\delta$ deep and $a$ wide

section defined by the skin depth and the width $a$ of the conductor slab. Then the resistance of the slab is given by

$$R = \frac{1}{\sigma}\frac{a}{a\delta} = \frac{1}{\sigma}\sqrt{\frac{\omega\mu\sigma}{2}} = \sqrt{\frac{\omega\mu}{2\sigma}},$$

which can be obtained in an exact manner using the complex Poynting's vector.

Equation (20.9) is used to determine the attenuation in all metallic systems for guiding electromagnetic energy, like two-wire lines, coaxial lines, and rectangular waveguides. We illustrate this with two examples.

**Example 20.4—Resistance and internal inductance per unit length of a cylindrical wire at high frequencies.** Consider a straight round wire of radius $a$, conductivity $\sigma$, and permeability $\mu$, carrying a sinusoidal current of angular frequency $\omega$ and with an rms value $I$. The magnetic field intensity on the wire surface is $H(0) = I/(2\pi a)$, so the Joule's losses per unit length of the wire, according to Eq. (20.9), are

$$P_J' = R_s \frac{I^2}{(2\pi a)^2} 2\pi a = R_s \frac{I^2}{2\pi a}.$$

Because the resistance per unit length is defined by the relation $P_J' = R'I^2$, we obtain that, at high frequencies,

$$R' = \frac{R_s}{2\pi a} \quad (\Omega/\text{m}). \tag{20.11}$$

*(Resistance per unit length of round conductor at high frequencies)*

According to Eq. (20.9), the reactive power at high frequencies inside the conductor per unit area is the same as the power of Joule's losses. We know from circuit theory that the internal reactive power per unit length of the wire can be expressed as $X_{int}'I^2$. The power in Eq. (20.9) refers to the field *inside the conductor* (i.e., the wire) *only*. Since it is positive, the internal reactance is inductive, that is, $X_{int}' = R' = \omega L_{int}'$. Therefore the *internal inductance* of the

wire at high frequencies per unit length is given by

$$L'_{\text{int}} = \frac{R'}{\omega} = \frac{R_s}{2\pi a\omega} \quad (\text{H}/\text{m}). \tag{20.12}$$

*(Internal inductance per unit length of round conductor at high frequencies)*

This formula for $L'_{\text{int}}$ was given in Table 18.1.

**Example 20.5—Resistance and internal inductance per unit length of a thin two-wire line at high frequencies.** Using the results from Example 20.4, it is a simple matter to calculate the resistance and internal inductance per unit length of a thin two-wire line. Let the line have conductors of radius $a$, and let the distance between the wires be much larger than $a$. Then the influence of the current in one wire on the current distribution inside the other can be neglected. That means that the current distribution in each wire is practically axially symmetric, as for a single wire in Example 20.4. Therefore, the resistance and internal inductance per unit length are just twice those calculated in the preceding example (because we have two wires). This is the formula given in Table 18.1.

*Questions and problems:*    Q20.11 and Q20.12, P20.1 to P20.12

## 20.3   Proximity Effect

The term *proximity effect* refers to the influence of alternating current in one conductor on the current distribution in another, nearby conductor. Qualitatively, it can also be explained by Lentz's law.

Consider a coaxial cable of finite length. Assume for the moment that there is an alternating current only in the inner conductor (for example, that it is connected to a generator), and that the outer conductor is not connected to anything. If the outer conductor is much thicker than the skin depth, there is practically no magnetic field inside the outer conductor. If we apply Ampère's law to a coaxial circular contour contained in that conductor, it follows that the induced current on the *inside* surface of the outer conductor is exactly equal and opposite to the current in the inner conductor. This is an example of the proximity effect.

The current from the inner surface of the outer conductor must close into itself over the *outside* surface of the outer conductor, so that on that surface the same current exists as in the inner conductor.

Let us now, in addition, have normal cable current in the outer conductor. It is the same, but opposite, to the current on the conductor outer surface, so the two cancel out. We are left with a current over the inner conductor, and a current over the inside surface of the outer conductor. This is a combined skin effect and proximity effect. Normally, this *combined* effect is what is actually encountered, but it is usually called just the proximity effect.

If the skin effect is not pronounced, the situation is similar except that there is an appreciable current density at all points of the inner and outer cable conductors, as sketched in Fig. 20.3.

The analysis of the proximity effect (i.e., of the combined proximity effect and skin effect) is rather complicated. We shall not, therefore, illustrate the proximity ef-
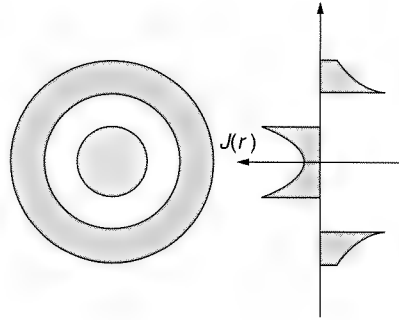
Figure 20.3 Combined proximity and skin effects in a coaxial cable

fect for the general case. If the skin effect is very pronounced, however, in some cases it becomes quite simple, as the next example shows.

**Example 20.6—Resistance and internal inductance per unit length of a coaxial cable at high frequencies.** Consider again a coaxial cable, with an inner conductor of radius $a$ and an outer conductor of inner radius $b$. Assume that a sinusoidal current of rms value $I$ flows through the cable at a frequency for which the skin effect is very pronounced. In this case, we have two thin current layers: one over the inner conductor, and one over the inside surface of the outer conductor, as explained previously. According to Eq. (20.9), Joule's losses per unit length of the cable (in both conductors) are given by the sum of losses in the cylinders of radius $a$ and of radius $b$. Therefore, the resistance per unit length is the sum of that in Eq. (20.11) and of the same expression with $a$ substituted by $b$:

$$R' = \frac{R_s}{2\pi a} + \frac{R_s}{2\pi b} = \frac{R_s}{2\pi}\left(\frac{1}{a} + \frac{1}{b}\right). \tag{20.13}$$

We know that the internal reactance per unit length has the same value as $R'$. The internal inductance of the cable at high frequencies is therefore $L'_{\text{int}} = R'/\omega$. These are the formulas given earlier in Table 18.1 without explanation.

## 20.4  Chapter Summary

1. Sinusoidal currents in good conductors are not distributed uniformly over their cross section. Rather, as frequency increases, the current tends to concentrate near the conductor surfaces, a phenomenon known as the *skin effect*.

2. At very high frequencies, the skin effect is so pronounced that current exists only over a very thin layer of any good (metallic) conductor.

3. The penetration of current in a good conductor is characterized by the *skin depth*. At this depth, the current density is $1/e \simeq 0.37$ of that at the conductor surface. At 60 Hz, it is on the order of 1 cm for copper and 1 mm for iron.

4. The skin depth is inversely proportional to the square root of frequency, permeability, and conductivity.

**5.** A time-varying current in one conductor influences the current distribution in nearby conductors, a phenomenon known as the proximity effect.

**6.** Both skin effect and proximity effect are consequences of electromagnetic induction.

**7.** In conducting magnetic materials, the time-varying magnetic field also exhibits the skin effect. For this reason, ferromagnetic cores of alternating-current machinery are made of thin, mutually insulated sheets. At very high frequencies, transformer and inductor cores are made of ferrites, which have a relatively high permeability, but are also relatively good insulators, so that the skin effect for the magnetic field almost does not exist.

# QUESTIONS

**Q20.1.** Three long parallel wires a distance $d$ apart are in one plane. At their ends they are connected together. These common ends are then connected by a large loop to a generator of sinusoidal emf. Are the currents in the three wires the same? Explain. [*Hint:* Have in mind Eq. (14.3), where $\mathbf{J}\,dv$ is substituted by $i\,d\mathbf{l}$.]

**Q20.2.** $N$ long parallel thin wires are arranged uniformly around a circular cylinder. At their ends the wires are connected by a large loop to a generator of sinusoidal emf. Are the currents in the $N$ wires the same? Explain.

**Q20.3.** Another wire is added in question Q20.2 along the axis of the cylinder. Is the current in the added wire the same as in the rest? Is it smaller or greater? Explain, having in mind Eq. (14.3).

**Q20.4.** A thin metallic strip of width $d$ carries a sinusoidal current of a high frequency. What do you expect the distribution of current in the strip to be like?

**Q20.5.** The two conductors of a coaxial line are connected in parallel to a generator of sinusoidal emf. Is the current intensity in the two conductors the same? If it is not, does the difference depend on frequency? Explain.

**Q20.6.** A metal coin is situated in a time-harmonic uniform magnetic field, with faces normal to the field lines. What are the lines of eddy currents in the coin like? What are the lines of the induced electric field of these currents?

**Q20.7.** So-called induction furnaces are used for melting iron by producing large eddy currents in iron pieces. Assume that the iron in the furnace is first in the form of small ferromagnetic objects (nails, screws, etc.). What do you expect to happen if they are exposed to a very strong time-harmonic magnetic field? What happens once they melt?

**Q20.8.** Two parallel, coplanar thin strips carry equal time-harmonic currents. What do you think the current distribution in the strips is like if the currents in the strips are (1) in the same direction, and (2) in opposite directions?

**Q20.9.** A thick copper conductor of square cross section carries a large time-harmonic current. Where do you expect the most intense Joule's heating of the conductor? Explain.

**Q20.10.** A ferromagnetic core of a solenoid is made of thin sheets. If the current in the solenoid is time-harmonic, where do you expect the strongest heating of the core due to eddy currents?

**Q20.11.** Describe the procedure of determining the resistance and internal inductance per unit length of a stripline at high frequencies. Neglect edge effect.

**Q20.12.** When compared with current density on the surface, what is the magnitude of current density in a thick conducting sheet one skin depth below the surface, and what is it at two skin depths below the surface?

# PROBLEMS

**P20.1.** Check all skin depth values given in Table 20.1.

**P20.2.** Starting from Eq. (20.7), prove that the total current in the half-space in Fig. 20.1 is the same as if a current of constant density $J_z(0)/(1+j)$ exists in a slab $0 \le y \le \delta$.

**P20.3.** Determine the total Joule's losses per unit area of the half-space in Fig. 20.1 by integrating the density of Joule's losses. Compare the result with Eq. (20.9).

**P20.4.** Using Poynting's theorem in complex form, prove that for any conductor with two close terminals, at very high frequencies the conductor resistance and internal reactance are equal. Find the (integral) expression for these quantities.

**P20.5.** A stripline of strip width $a = 2\,\mathrm{cm}$, distance between them $d = 2\,\mathrm{mm}$, and the thickness of the strips $b = 1\,\mathrm{mm}$ carries a time-harmonic current of rms value $I = 0.5\,\mathrm{A}$ and frequency $f = 1\,\mathrm{GHz}$. The strips are made of copper. Neglecting fringing effect, determine the line resistance and total inductance per unit length.

**P20.6.** Starting from Eqs. (20.3), determine the distribution of current in a flat conducting sheet of thickness $d$. The sheet conductivity is $\sigma$, permeability $\mu$, and angular frequency of the current is $\omega$. Set the origin of the $y$ coordinate at the sheet center, and assume that the rms value of the current density at the center is $J_z(0)$. Plot the resulting current distribution.

**\*P20.7.** Find $H_x(y)$ from Eqs. (20.3) and (20.7), and $E_z(y)$ from Eq. (20.7). Use these expressions and Poynting's theorem to prove Eq. (20.9).

**P20.8.** Starting from Eq. (20.7), derive the expression for the instantaneous value of the current density, $J_z(y, t)$.

**P20.9.** Calculate the resistance per unit length of a round copper wire of radius $a = 1\,\mathrm{mm}$, from the frequency for which the skin depth is one-tenth of the wire radius, to the frequency $f = 10\,\mathrm{GHz}$. Plot this resistance as a function of frequency.

**P20.10.** Assume that in a ferromagnetic round wire of radius $a$, conductivity $\sigma$, and permeability $\mu$, there is an axial magnetic field of angular frequency $\omega$ and of rms flux density $B$ practically constant over the wire cross section. Find the expressions for eddy currents in the wire and eddy current losses in the wire per unit length.

**P20.11.** A bunch of $N$ insulated round wires of radius $a$, conductivity $\sigma$, and permeability $\mu$ is exposed to an axial time-harmonic magnetic field of angular frequency $\omega$. The frequency is sufficiently low that the field can be considered uniform over the cross section of the wires. If the rms value of the magnetic flux density is $B_0$, determine the time-average eddy current power losses in the bunch, per unit volume *of the wires*. Use the result of the preceding problem. Specifically, calculate the losses per unit volume assuming $B_0 = 0.1\,\mathrm{T}$, $a = 0.5\,\mathrm{mm}$, $\sigma = 10^7\,\mathrm{S/m}$, $\mu = 1000\mu_0$, and $f = 60\,\mathrm{Hz}$.

**\*P20.12.** Consider a straight wire of radius $a$, conductivity $\sigma$, and permeability $\mu$. Let the wire axis be the $z$ axis of a cylindrical coordinate system. Assume there is a current in the wire of rms value $I$ and angular frequency $\omega$. Starting from Maxwell's equations in cylindrical coordinates, derive the differential equation for the only existing, $J_z$ component of the current density vector in the wire. Note that, by symmetry, the only component of **H** is $H_\phi$. Do *not* attempt to solve the equation you obtain. (If your equation is correct, it is known as a Bessel differential equation, and its solutions are known as Bessel functions.)

# 21

# *Uniform Plane Waves*

## 21.1   Introduction

Maxwell's theory predicts the existence of specific electromagnetic fields, known as *electromagnetic waves*. These fields, once created by time-varying currents and charges, continue to move with a finite velocity independent of the sources that produced them.

Much of the rest of this book is devoted to the analysis of various types of electromagnetic waves. In this chapter the simplest type of wave, known as a *uniform plane wave*, is considered. Although they are the simplest, uniform plane waves are of extreme practical importance: actual waves radiating from sources are spherical, but at large distances from sources they become practically plane waves; and in addition, more complicated wave types can be represented as a superposition of plane waves.

## 21.2   The Wave Equation

The wave equation is a second-order partial differential equation that is satisfied by all electromagnetic fields in *homogeneous linear media*.

Assume that an electromagnetic field exists in a homogeneous linear medium with parameters $\epsilon$, $\mu$, and $\sigma$. Suppose that there are neither free charges nor field sources (impressed electric fields) in the medium considered. Maxwell's equations in

that case have the form

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \qquad \nabla \cdot \mathbf{E} = 0, \tag{21.1}$$

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t}, \qquad \nabla \cdot \mathbf{H} = 0. \tag{21.2}$$

To eliminate $\mathbf{H}$ from the first equation pair, let us apply the curl operator to the first equation. Since the curl implies differentiation with respect to space coordinates, it is independent of the differentiation with respect to time. Therefore, $\nabla \times (\partial \mathbf{H}/\partial t)$ can be written as $\partial(\nabla \times \mathbf{H})/\partial t$. With this in mind, making use of the first of Eqs. (21.2), the first of Eqs. (21.1) becomes

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu\sigma \frac{\partial \mathbf{E}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2}. \tag{21.3}$$

In a similar manner we eliminate $\mathbf{E}$ from the first equation of the second pair, to obtain

$$\nabla \times (\nabla \times \mathbf{H}) = -\mu\sigma \frac{\partial \mathbf{H}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2}. \tag{21.4}$$

We know from vector analysis that for any vector function $\mathbf{F}$ that can be differentiated twice, $\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2 \mathbf{F}$ [see Appendix 2, No. 28]. According to the second parts of Eqs. (21.1) and (21.2), $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{H} = 0$, so Eqs. (21.3) and (21.4) become

$$\nabla^2 \mathbf{E} - \epsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} - \mu\sigma \frac{\partial \mathbf{E}}{\partial t} = 0, \tag{21.5}$$

*(Wave equation for vector $\mathbf{E}$)*

and

$$\nabla^2 \mathbf{H} - \epsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2} - \mu\sigma \frac{\partial \mathbf{H}}{\partial t} = 0. \tag{21.6}$$

*(Wave equation for vector $\mathbf{H}$)*

These are the *wave equations.*

If the field is time-harmonic and we use complex notation, we obtain

$$\nabla^2 \underline{\mathbf{E}} + (\omega^2 \epsilon\mu - j\omega\mu\sigma)\underline{\mathbf{E}} = 0, \tag{21.7}$$

*[Helmholtz equation (complex wave equation) for vector $\mathbf{E}$]*

and

$$\nabla^2 \underline{\mathbf{H}} + (\omega^2 \epsilon\mu - j\omega\mu\sigma)\underline{\mathbf{H}} = 0. \tag{21.8}$$

*[Helmholtz equation (complex wave equation) for vector $\mathbf{H}$]*

Although these are just wave equations in complex form, sometimes they are referred to as the *Helmholtz equations.*

Note that the wave equations were derived by using the conditions $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{H} = 0$, but these conditions are *not* implicit in the wave equations. Because they must be satisfied, a solution must be sought using the following procedure:

1. We find a solution to the wave equation, for example, Eq. (21.5), that satisfies the condition $\nabla \cdot \mathbf{E} = 0$.

2. We determine vector $\mathbf{H}$ from the first of Eqs. (21.1).

The second step guarantees that also $\nabla \cdot \mathbf{H} = 0$ (recall that the divergence of the curl is zero). The first of Eqs. (21.2) is thereby also satisfied: the time derivative of that equation, with $\partial \mathbf{H}/\partial t$ from the first of Eqs. (21.1), becomes the wave equation for $\mathbf{E}$. Consequently, the solution found in this way satisfies all the necessary equations and is a legitimate solution of Maxwell's equations.

Example 21.1—Wave equation in a rectangular coordinate system. The wave equations (21.5) to (21.8) are valid in any coordinate system. What do they become in a rectangular coordinate system?

Note first that these equations are *vector equations*, that is, they represent *three scalar equations*. Consider, for example, the wave equation for vector $\mathbf{E}$, Eq. (21.5). Recalling the expression for $\nabla^2 \mathbf{E}$ in a rectangular coordinate system, we obtain for the $x$ component of the wave equation

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} - \epsilon\mu\frac{\partial^2 E_x}{\partial t^2} - \mu\sigma\frac{\partial E_x}{\partial t} = 0. \tag{21.9}$$

The $y$ and $z$ components of the equation are of exactly the same form, with $E_x$ replaced by $E_y$ and $E_z$, respectively.

*Questions and problems:*   Q21.1 and Q21.2

## 21.3   Uniform Plane Electromagnetic Waves in Perfect Dielectrics

We now use the wave equation to analyze a specific electromagnetic field. Let the field satisfy the following two conditions:

1. Both vectors $\mathbf{E}$ and $\mathbf{H}$ depend only on coordinate $z$ and time.

2. The field exists in a homogeneous, lossless medium, with parameters $\epsilon$, $\mu$, and $\sigma = 0$.

We first stipulate that $\nabla \cdot \mathbf{E} = 0$. Because in a rectangular coordinate system

$$\nabla \cdot \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z},$$

and because we assumed $\mathbf{E}$ to depend only on $z$ and time, the first two terms on the right-hand side are zero. So the condition $\nabla \cdot \mathbf{E} = 0$ reduces to $\partial E_z/\partial z = 0$. This is satisfied if $E_z = 0$, or possibly $E_z = $ constant. The constant solution (with respect to $z$) is not of interest because we search for a field varying along the $z$ axis, so $E_z = 0$.

No generality is lost if we assume that only one of the two components normal to $z$, $E_x$ or $E_y$, is nonzero. Let $E_x$ be nonzero, and $E_y = 0$. Then the wave equation

for the vector **E** has only an $x$ component, so that Eq. (21.9) is the only one existing. Because the derivatives with respect to $x$ and $y$ are zero (field components depend on the coordinate $z$ only), and $\sigma = 0$ for a perfect dielectric, Eq. (21.9) becomes

$$\frac{\partial^2 E_x}{\partial z^2} - \epsilon \mu \frac{\partial^2 E_x}{\partial t^2} = 0. \tag{21.10}$$

This equation has the same form as Eqs. (18.5) that we derived for the voltage and current along a transmission line. Notice that instead of $L'C'$ in Eqs. (18.5), we now have $\epsilon\mu$, and we know these products are equal for lossless lines. As we have seen in Chapter 19, the solution to this equation is of the form

$$E_x(z, t) = E_1 f_1 \left( t - \frac{z}{c} \right) + E_2 f_2 \left( t + \frac{z}{c} \right). \tag{21.11}$$

*(E field consisting of incident and reflected plane waves)*

In this equation,

$$c = \frac{1}{\sqrt{\epsilon\mu}} \tag{21.12}$$

*(Velocity of propagation of plane waves)*

is the velocity of propagation of plane waves, $E_1$ and $E_2$ are constants, and $f_1$ and $f_2$ are *any* functions of the arguments $(t - z/c)$ and $(t + z/c)$, respectively. That the expression in Eq. (21.11) is the solution of Eq. (21.10) can be proved by substitution, as we did in Chapter 18. In fact, there are an infinite number of solutions to the wave equation (infinite number of different fields), since $f_1(t - z/c)$ and $f_2(t + z/c)$ are *arbitrary* functions.

**Example 21.2—Some specific wave functions.** Let us construct a few specific wave functions. For this, we need to consider *any* function of a single variable, and to replace this variable by $(t \pm z/c)$. For example, consider the function $\sin \omega t$. The corresponding wave function is $\sin \omega(t \pm z/c)$. A wave function $e^{\pm j\omega(t \pm z/c)}$ corresponds to the function $e^{\pm j\omega t}$. As a final example, consider the following function:

$$f(x) = 1 \quad \text{for} \quad a < x < b \quad \text{else} \quad f(x) = 0.$$

The corresponding wave function, for example of the form $f(t - z/c)$, is

$$f(t - z/c) = 1 \quad \text{for} \quad a < (t - z/c) < b \quad \text{else} \quad f(t - z/c) = 0.$$

We know from Chapter 18 what the physical meaning of the functions $f_1(t - z/c)$ and $f_2(t + z/c)$ is: $f_1(t - z/c)$ is an incident (forward) traveling (electric field) wave, and $f_2(t + z/c)$ is a reflected (backward) traveling wave (with respect to the $z$ axis). This important conclusion for $f_1(t - z/c)$ is illustrated again in Fig. 21.1. Such moving fields, as already mentioned, are known as *electromagnetic waves*. Note again that the
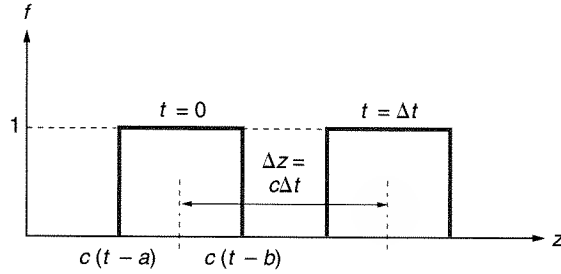
Figure 21.1 A pulse at $t = 0$ and at a later instant $\Delta t$, as described in Example 21.2

$E$-field component satisfies the same equation and that this equation has the same solutions as that for the voltage along a transmission line.

We have thus determined the electric field of a wave. Let us now determine the magnetic field of the wave, to obtain a complete solution to Maxwell's equations. This is analogous to determining the current along a transmission line from the transmission-line equations, and then finding the characteristic impedance as a ratio of voltage and current at a point. Again, the analogy is not surprising, as the magnetic field is produced by the current.

To simplify the derivation, consider only the wave with the argument $(t - z/c)$. We proceed as outlined in the preceding section and determine $\mathbf{H}$ from the first of Eqs. (21.1). Having in mind the expression for the curl in a rectangular coordinate system, and that $\mathbf{E}(z, t) = E_x(z, t)\mathbf{u}_x$, this *vector* equation results in the following three scalar equations:

$$0 = -\mu\frac{\partial H_x}{\partial t} \qquad \frac{\partial E_x}{\partial z} = E_1\frac{\partial f_1}{\partial z} = -\mu\frac{\partial H_y}{\partial t} \qquad 0 = -\mu\frac{\partial H_z}{\partial t}. \qquad (21.13)$$

We are not interested in time-constant field components (the components having zero time derivative), so $H_x = H_z = 0$. The only existing time-varying component of the magnetic field intensity, $H_y$, is obtained by integrating the second of Eqs. (21.13). Having in mind Eq. (18.9), we obtain

$$H_y(z, t) = \frac{1}{\mu c}E_1 f_1\left(t - \frac{z}{c}\right) = \frac{1}{\mu c}E_x(z, t), \qquad (21.14)$$

or, since $c = 1/\sqrt{\epsilon\mu}$,

$$H_y(z, t) = \sqrt{\frac{\epsilon}{\mu}}E_x(z, t). \qquad (21.15)$$

($H_y$ component associated with the $E_x$ component of incident plane wave)

An interesting conclusion follows from this relation: for a plane wave, the energy density of the magnetic field, $\mu H^2/2$, at all points and at all instants, equals the energy

density of the electric field, $\epsilon E^2/2$. Of course, this is true only for a single incident plane wave, not for a possible superposition of waves.

The electric and magnetic field vectors of the electromagnetic wave we considered are in planes perpendicular to the direction of propagation of the wave (the $z$ direction). This is why the wave is known as a *plane wave*. In addition, these vectors are constant at any of these planes at a given instant. For this reason it is said that this particular plane wave is *uniform*. (Some plane waves are not uniform—for example, the electric and magnetic field waves along transmission lines are plane, but not uniform. Why?) Finally, since the vectors **E** and **H** are transverse to the direction of propagation, this kind of wave is known as a *transverse electromagnetic wave*, or *TEM wave*.

It is a simple matter to show that the unit for $\sqrt{\mu/\epsilon}$ is the ohm, that is, this quantity has the dimension of impedance. For this reason it is known as the *intrinsic impedance* (or sometimes just *impedance*) of the medium in which the wave exists. Note that Eq. (21.15) is analogous to Eq. (18.19), where the characteristic impedance of a transmission line was given by $\sqrt{L'/C'}$. The intrinsic impedance of a plane wave is usually denoted by $\eta$ (or sometimes $Z$),

$$\eta = \sqrt{\frac{\mu}{\epsilon}} \quad (\Omega). \tag{21.16}$$

*(Intrinsic impedance of the medium)*

The most important (and frequent) waves in practice are those propagating in a vacuum (or air). In that case, the intrinsic impedance of the medium and the velocity of propagation become

$$\eta_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \simeq 120\pi \ \Omega \simeq 377 \ \Omega \tag{21.17}$$

*(Intrinsic impedance of a vacuum)*

$$c_0 = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \simeq 3 \cdot 10^8 \ \text{m/s}. \tag{21.18}$$

*(Velocity of propagation of plane waves in a vacuum)*

Thus, *the velocity of propagation of plane waves in a vacuum is equal to the velocity of light in a vacuum.* This fact was the basis on which Maxwell argued that light is nothing but a rapidly oscillating electromagnetic wave. This mathematically obtained number was the basis for Maxwell's electromagnetic theory of light.

Note that the cross product of the vectors **E** and **H**, that is, the Poynting vector, is in the direction of propagation of the wave:

$$\mathbf{E} \times \mathbf{H} = E_x(z, t)H_y(z, t)\mathbf{u}_z = \mathcal{P}(z, t)\mathbf{u}_z. \tag{21.19}$$

This means that the wave *transports electromagnetic energy.*

If the $E$ field has a $y$ component (instead of an $x$ component), the Poynting theorem tells us that the $H$ field will have a $-x$ component, since the cross product $\mathbf{E} \times \mathbf{H}$ must be in the $+z$ direction (the direction of propagation of the wave). The relation between the two is

$$H_x(z, t) = -\sqrt{\frac{\epsilon}{\mu}} E_y(z, t). \tag{21.20}$$

This conclusion is important for understanding reflections of plane waves.

**Example 21.3—Plane waves propagating in the $-z$ direction.** Consider now the other solution of the wave equation, $E_x(z, t) = E_2 f_2(t + z/c)$. To determine the corresponding vector **H**, we just replace $c$ by $-c$ in the preceding derivations. So for the "reflected" wave

$$H_y(z, t) = -\sqrt{\frac{\epsilon}{\mu}} E_x(z, t),$$

and

$$H_x(z, t) = \sqrt{\frac{\epsilon}{\mu}} E_y(z, t).$$

It is left as an exercise for the reader to prove that the Poynting vector in both of these cases is directed in the $-z$ direction.

Let us summarize the properties of a uniform plane electromagnetic wave:

1. The vectors **E** and **H** are mutually perpendicular, and perpendicular to the direction of wave propagation.
2. The direction of the Poynting vector is in the direction of wave propagation.
3. At any instant, the magnitudes of vectors **E** and **H** are the same in any individual plane normal to the direction of propagation.
4. For a single plane wave (but not for a wave obtained as a superposition of several waves propagating in arbitrary directions), the ratio of the magnitudes of vectors **E** and **H** is the same at all points and at all instants. It is equal to the intrinsic impedance of the medium in which the wave exists, $\eta = \sqrt{\mu/\epsilon}$. In air and vacuum, $\eta_0 \simeq 120\pi \ \Omega \simeq 377 \ \Omega$.
5. The velocity of propagation of plane waves is $c = 1/\sqrt{\epsilon\mu}$; in a vacuum, this is equal to the velocity of light in a vacuum, $c_0 \simeq 3 \cdot 10^8 \ \text{m/s}$.

*Questions and problems:*   Q21.3 and Q21.4, P21.1 and P21.2

## 21.4    Time-Harmonic Uniform Plane Waves and Their Complex Form

Electromagnetic waves in electrical engineering are most often harmonic in time, or at least harmonic during a certain time interval. Suppose a wave at a fixed coordinate $z$ varies in time as $\cos \omega t$. In that case, for a wave propagating in the $+z$ direction we have

$$
E_x(z, t) = E\sqrt{2} \cos \omega(t - z/c) \qquad H_y(z, t) = \sqrt{\frac{\epsilon_0}{\mu_0}} E_x(z, t), \qquad (21.21)
$$
*(Electric and magnetic fields of a sinusoidal plane wave)*

where $E$ is the rms value of the electric field.

We know that $\cos(\alpha + n \cdot 2\pi) = \cos \alpha$ for all integer values of $n$. This means that the values of the electric and magnetic fields periodically repeat themselves in *time and space*. Namely, for a given $z = z_0$, the fields are changing according to $\cos \omega(t - z_0/c)$. On the other hand, for a given $t = t_0$, the fields change along the $z$ axis according to

$$
E_x(z, t_0) = E\sqrt{2} \cos(\omega t_0 - \omega z/c). \qquad (21.22)
$$

$E_x(z, t_0)$ has the same value at all points for which

$$
|\omega t_0 - \omega z_n/c| = n \cdot 2\pi \qquad n = 0, 1, 2, \ldots. \qquad (21.23)
$$

The distance along the $z$ axis between two such points $z_{n+1}$ and $z_n$ is

$$
\lambda = \frac{2\pi c}{\omega}. \qquad (21.24)
$$

We know from Chapter 18 that this distance is called the *wavelength* of the time-harmonic (sinusoidal) plane wave. Since $\omega = 2\pi f$, where $f$ is the frequency of the wave, the wavelength can also be expressed as

$$
\lambda = \frac{c}{f} \qquad (\text{m}), \qquad (21.25)
$$
*(Wavelength of plane waves)*

which we already know from Chapter 18.

Figure 21.2 shows the way an electromagnetic field changes in space (along the $z$ axis) when frozen in time. In time, the whole picture moves in the direction of the $z$ axis with a velocity $c$.

Time-harmonic electromagnetic waves used in engineering and physics have frequencies in a very wide range—from about 1 Hz to about $10^{22}$ Hz. This corre-
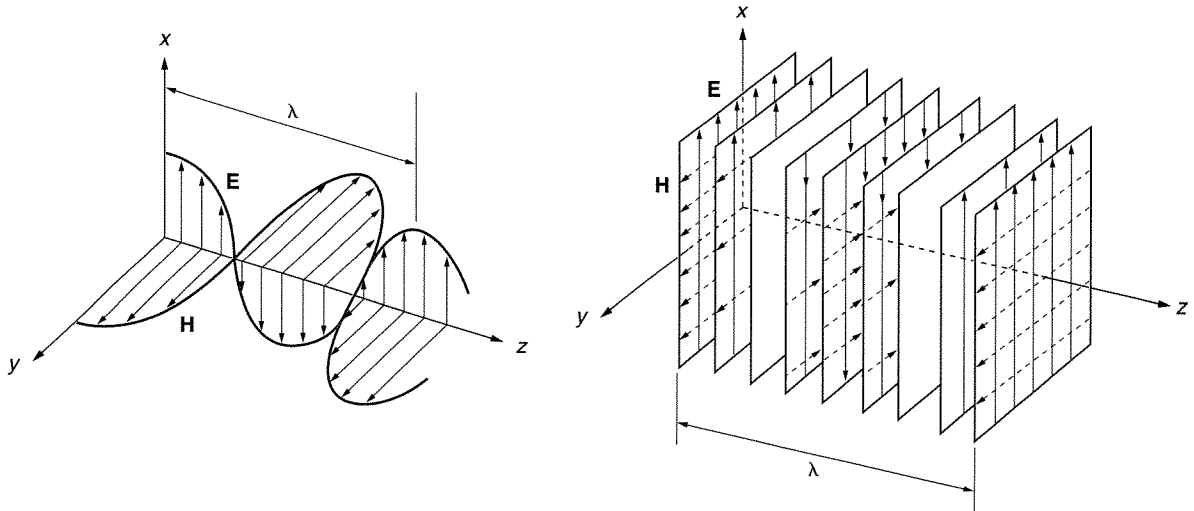
Figure 21.2 Two schematic ways of representing a time-harmonic incident plane wave frozen in time. As time passes, the picture moves in the +z direction with a velocity c.

sponds to wavelengths from about $10^8$ m to about $10^{-14}$ m. This range of frequencies is known as the *electromagnetic spectrum*. It is sketched in Fig. 21.3.

We can always determine the wavelength of a plane wave of a given frequency (or the frequency of a wave of a given wavelength) by means of the formula in Eq. (21.25). Electronics engineers usually use frequency, optical engineers use wavelength, and both are used in the microwave frequency range (from about 1 to 300 GHz). For the radio frequency (rf) and microwave region and for waves in a vacuum, the following simple rule is convenient to use: the wavelength in centimeters is equal to 30 divided by the frequency in GHz. For example, the wavelength at 10 GHz is 3 cm, and the wavelength at 900 MHz $= 0.9$ GHz is $33\frac{1}{3}$ cm.

Time-harmonic electromagnetic waves can be represented in complex form, as in the case of time-harmonic currents and voltages. We know that in the case of cur-
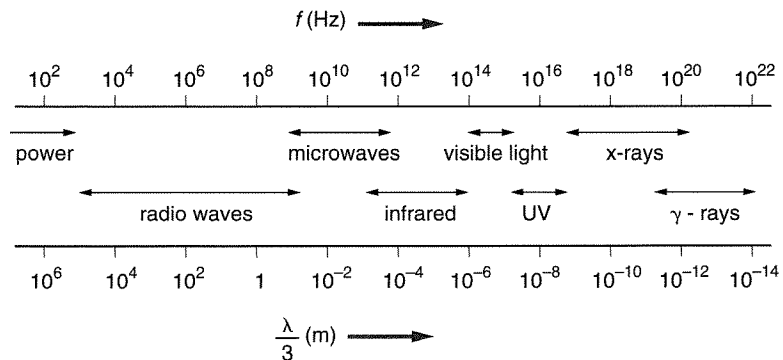


Figure 21.3 The electromagnetic spectrum

rents and voltages their form in time domain is obtained by multiplying their complex form by $\sqrt{2}e^{j\omega t}$, and taking the real part of the expression obtained in this way.

Let us write the expression in Eq. (21.21) for the $E$ field of a plane wave as

$$E_x(z, t) = E\sqrt{2}\cos(\omega t - \beta z + \theta), \tag{21.26}$$

where

$$\beta = \frac{\omega}{c} = \frac{2\pi}{\lambda} \qquad \text{(radian/m)}, \tag{21.27}$$

*(Phase coefficient for plane waves)*

as earlier, is termed the *phase coefficient*, and $\theta$ is the initial phase of the field [which in Eq. (21.21) was assumed to be zero]. (Because for various media in which plane waves may propagate the phase coefficient may *not* be constant, the term "phase constant" for waves is not quite appropriate.) The complex form of the expression in Eq. (21.26) should be such that we obtain Eq. (21.26) from it in the same way we obtain time-domain forms of currents and voltages from their complex forms:

$$E_x(z, t) = \sqrt{2}\operatorname{Re}\left\{\underline{E}_x(z)e^{j\omega t}\right\}. \tag{21.28}$$

By comparing the last expression with that in Eq. (21.26), we identify immediately the complex form of the $E$ field (and $H$ field) of a plane wave propagating along the $z$ axis:

$$\underline{E}_x(z) = \underline{E}e^{-j\beta z} \qquad \underline{H}_y(z) = \frac{1}{\eta}\underline{E}e^{-j\beta z} \qquad \text{where} \qquad \underline{E} = Ee^{j\theta}. \tag{21.29}$$

Thus, as we already know from transmission lines, a factor $e^{-j\beta z}$ in the complex form of a quantity indicates propagation of that quantity in the $+z$ direction with a velocity $c = \omega/\beta$. Evidently, a factor of the form $e^{+j\beta z}$ indicates propagation in the $-z$ direction.

*Questions and problems:*   Q21.5 to Q21.11, P21.3 to P21.10

## 21.5    Polarization of Plane Waves

Any number of plane waves propagating in the same direction add up to a complex plane wave. The time variation of the component waves may be arbitrary. Fortunately, such a general case is of quite minor engineering interest. We are mostly interested in time-harmonic plane waves. For these simple uniform plane waves, the vector **E** is at all times parallel to an axis (the $x$ axis in our case). It does vary in time, but the tip of the vector traces a line parallel to the $x$ axis. We say that the polarization of this wave is *linear*.

What happens if we have two otherwise arbitrary plane waves, whose frequencies are the same, propagating in the same direction? We shall now show that the resulting wave is a plane wave with the tip of vector **E** tracing an ellipse at every point in space. We say that such a resulting wave is *elliptically polarized*. In the special case when the major and minor semi-axes of the ellipse are of equal lengths, the tip of vector **E** at a fixed point in space traces a circle. We say that the polarization of this wave is *circular*.

To demonstrate elliptic polarization of plane waves, consider two plane waves of the same frequency and propagating in the same direction, but with vectors **E** of the two waves perpendicular to each other and of different amplitudes. Let the directions of the two $E$ vectors be the $x$ and $y$ direction, and let one of them vary as the cosine function, and the other as the sine function:

$$E_x(z, t) = E_1 \cos(\omega t - \beta z), \tag{21.30}$$

$$E_y(z, t) = E_2 \sin(\omega t - \beta z). \tag{21.31}$$

In this simple case

$$\left[ \frac{E_x(z, t)}{E_1} \right]^2 + \left[ \frac{E_y(z, t)}{E_2} \right]^2 = 1, \tag{21.32}$$

since $\cos^2 \alpha + \sin^2 \alpha = 1$.

For a fixed $z$, this is the equation of an *ellipse* with semi-axes $E_1$ and $E_2$. This means that indeed, the tip of the total electric field intensity vector, $\mathbf{E}_{\text{tot}}(z, t) = E_x(z, t)\mathbf{u}_x + E_y(z, t)\mathbf{u}_y$, for any fixed $z$, in the course of time describes this ellipse.

If $E_1 = E_2$, the tip of $\mathbf{E}_{\text{tot}}(z, t)$, for any fixed $z$, describes a circle of radius $E_1$, so the total field is circularly polarized.

Figure 21.4 illustrates elliptic, circular, and linear polarizations. A representation of an elliptically polarized incident plane wave frozen in time is shown in Fig. 21.5. In the figure, the density of the $E$ lines (solid) and $H$ lines (dashed) is proportional to the local intensity of these vectors. (Note the different density of lines along the $z$ axis.) The whole picture moves with a velocity $c_0$ in the $+z$ direction.
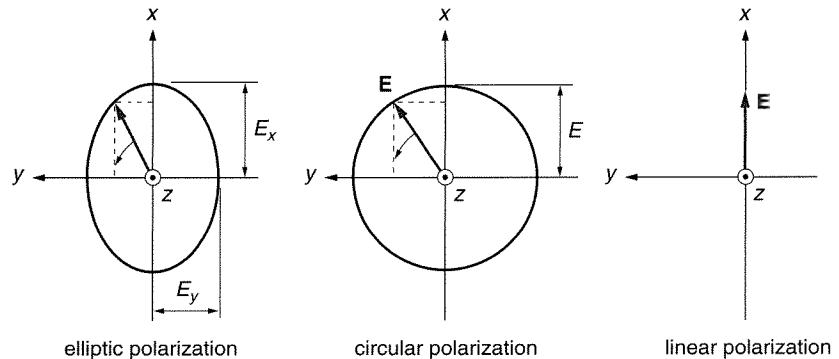


Figure 21.4 Illustration of elliptic, circular, and linear polarizations
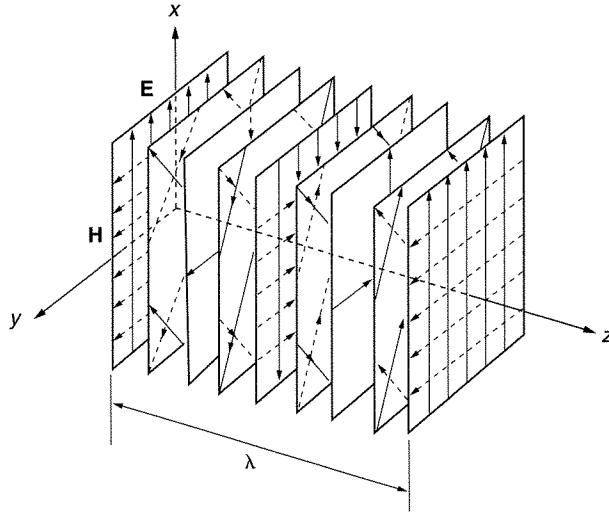
Figure 21.5 A representation of an elliptically polarized plane wave

So, *for a fixed z*, the vector **E** (and, of course, also the vector **H**) rotates and changes the magnitude, its tip describing an ellipse. This rotation can be in one or the other direction. If the direction of rotation is given by the right-hand rule with respect to the direction of wave propagation, it is said that the polarization of the wave is *right-handed*. If the field vectors rotate in the opposite direction, the polarization is said to be *left-handed*.

*Questions and problems:*    Q21.12 to Q21.17, P21.11 to P21.15

## 21.6   Phase Velocity and Group Velocity: Dispersion

We know that the velocity of propagation of uniform plane electromagnetic waves is $c = 1/\sqrt{\epsilon\mu}$. In a vacuum, this velocity equals the velocity of light, and is the same for any frequency of the wave. In other media, the permittivity and permeability depend on frequency at least to some extent. Therefore, except in a vacuum, the phase coefficient, and therefore also the velocity of propagation of uniform plane waves, depend on the wave frequency. This is known as *dispersion*.

In many media, dispersion can be ignored. In quite a number of important cases, however, it must be taken into account, as it results in signal distortion. Namely, any signal is composed of time-harmonic components contained in a certain frequency band. Its shape is determined by relative amplitudes *and phases* of the time-harmonic components in this band. If the velocities of the time-harmonic components are not the same, their relative positions, which means their relative phases, change as the signal propagates, which means that the signal shape changes

as well. For this reason, the frequency band of a signal is usually made small enough so that distortion can be ignored.

The velocity $c$ of plane waves in Eq. (21.21) is the velocity with which the *phase* of the wave propagates. To be more specific, it determines the progression of the $z$ coordinates in the argument of the cosine function, which ensures that as time passes, the argument (i.e., the phase) of the cosine function remains unchanged. For this reason, the velocity $c$ is termed the *phase velocity*. According to Eq. (21.27), the phase velocity, $v_{\text{ph}} = c$, can be expressed as

$$v_{\text{ph}}(\omega) = \frac{\omega}{\beta(\omega)} \quad \text{(m/s)}. \tag{21.33}$$

*(Definition of phase velocity)*

There is another important concept connected to dispersion, known as the *group velocity*. It represents the velocity of the signal in a dispersive medium and can be defined only for cases where dispersion is small.

To determine the group velocity, consider a simple signal in a weakly dispersive medium. Let the signal be obtained as a superposition of two plane waves propagating in the same direction and with slightly different angular frequencies, $\omega_1$ and $\omega_2$, and slightly different phase coefficients, $\beta_1$ and $\beta_2$ (due to dispersion). Without loss of generality, we can assume the amplitudes of both waves are the same, for example equal to 1, and consider a signal $f(z, t)$ of the form

$$f(z, t) = \cos(\omega_1 t - \beta_1 z) + \cos(\omega_2 t - \beta_2 z). \tag{21.34}$$

Now, $\cos a_1 + \cos a_2 = 2 \cos[(a_2 - a_1)/2] \cos[(a_2 + a_1)/2]$, so that

$$f(z, t) = 2 \cos(\Delta\omega\, t - \Delta\beta_1 z) \cos(\omega t - \beta z), \tag{21.35}$$

where

$$\Delta\beta = \frac{\beta_2 - \beta_1}{2}, \qquad \Delta\omega = \frac{\omega_2 - \omega_1}{2}, \qquad \omega = \frac{\omega_1 + \omega_2}{2}, \qquad \beta = \frac{\beta_1 + \beta_2}{2}. \tag{21.36}$$

Since $\Delta\omega \ll \omega$, the shape of this signal frozen in time is as sketched in Fig. 21.6. This is a rapidly varying wave (of frequency $\omega$) modulated by a slowly varying wave (of frequency $\Delta\omega$). The velocity of propagation of the rapidly varying modulated wave (the solid line in Fig. 21.6) is simply $\omega/\beta$. The velocity of the modulating wave (signal), however, is different—it is equal to the velocity of the *envelope* of the rapidly varying wave, indicated by the dashed lines. This means that the group velocity, $v_g$, is given by

$$v_g = \frac{\Delta\omega}{\Delta\beta} = \frac{1}{\Delta\beta/\Delta\omega}. \tag{21.37}$$

$$f_1(0, z) + f_2(0, z)$$
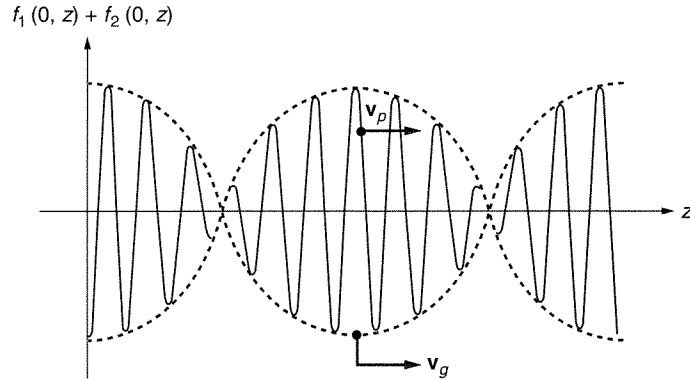


Figure 21.6 Sum of two time-harmonic functions of equal amplitudes and slightly different frequencies

What is the meaning of this expression? We know that $\beta$ is a function of $\omega$. For example, for a homogeneous dispersive dielectric of parameters $\epsilon(\omega)$ and $\mu(\omega)$ we know that $\beta(\omega) = \omega\sqrt{\epsilon(\omega)\mu(\omega)}$. Since $\Delta\beta$ is a small difference in $\beta$ corresponding to a small difference in $\omega$, the expression in the denominator in the last expression is in fact the derivative $d\beta(\omega)/d\omega$. We finally have

$$v_g(\omega) = \frac{1}{d\beta(\omega)/d\omega} \qquad \text{(group velocity).} \tag{21.38}$$

Because the envelope in Fig. 21.6 represents a signal (this is called *amplitude modulation*, or AM), the *group velocity* is the velocity of propagation of information transmitted by electromagnetic waves.

**Example 21.4—Group velocity in nondispersive media.** Assume that the medium is not dispersive. In that case, $\beta(\omega) = \omega/c$. The formula in Eq. (21.38) for the group velocity thus yields $v_g = c$. Of course, this was expected because in nondispersive media all time-harmonic components propagate with the same velocity, and therefore the group velocity is the same as the phase velocity.

**Example 21.5—Phase velocity and group velocity in ionized gases and hollow metal waveguides.** We will show in Chapters 23 and 25 that the phase coefficient for both a hollow metal waveguide and a homogeneous ionized gas is of the form

$$\beta(\omega) = \frac{\omega}{c_0}\sqrt{1 - \frac{\omega_c^2}{\omega^2}},$$

where $\omega_c$ is a constant that depends on the medium through which the wave propagates. The phase velocity is given by Eq. (21.33):

$$v_{\text{ph}}(\omega) = \frac{\omega}{\beta(\omega)} = \frac{c_0}{\sqrt{1 - \omega_c^2/\omega^2}} \quad \text{(m/s).} \tag{21.39}$$

The group velocity is obtained from Eq. (21.38). The final result is

$$v_g(\omega) = \frac{1}{d\beta(\omega)/d\omega} = c_0\sqrt{1 - \frac{\omega_c^2}{\omega^2}} \quad \text{(m/s)}.$$ (21.40)

So we see that the group velocity is less than $c_0$ (the velocity of light in a vacuum), but the phase velocity is larger than $c_0$! How this can be, when we know that $c_0$ is the largest possible velocity? Note that the phase velocity is a purely *geometrical* velocity, not the velocity of a particle or of a wave, so it can have any value, even larger than $c_0$. However, a wave does not transport power or information at the phase velocity, but rather at the group velocity, which is always smaller than $c_0$.

## 21.7 Chapter Summary

1. Maxwell's equations predict the existence of a specific type of electromagnetic field that, once created by time-varying currents and charges, continues to exist with no connection whatsoever with its sources. Such fields are known as *electromagnetic waves.*

2. The simplest electromagnetic waves are uniform plane waves. Their electric and magnetic field vectors are normal to each other and to the direction of propagation, and constant in planes normal to that direction. They are therefore known as *uniform transverse electromagnetic (TEM) waves.*

3. The speed of plane waves equals $1/\sqrt{\epsilon\mu}$, which in a vacuum equals the speed of light.

4. The ratio of the electric and magnetic fields at any point of a plane wave, and at all instants, is a constant, equal to the intrinsic impedance of the medium, $\eta = \sqrt{\mu/\epsilon}$.

5. If the properties of a medium depend on frequency, it is called a *dispersive medium.* For a signal composed of a narrow frequency band, and if dispersion is small, the signal propagates with a velocity known as the *group velocity.* The group velocity is different from the phase velocity, which is a geometrical velocity with which a fixed phase of the wave propagates.

6. The tip of the electric field vector of a time-harmonic field at a fixed point in space may trace in the course of time a straight-line segment (linear polarization), a circle (circular polarization) or an ellipse (elliptic polarization). No other traces are possible for a time-harmonic field of a single frequency.

7. The circular or elliptic polarizations are said to be right-handed if the rotation is clockwise, looking in the direction of the wave propagation. For waves rotating in the opposite direction, the polarization is said to be left-handed.

### QUESTIONS

**Q21.1.** What would Eqs. (21.1) and (21.2) be like for an inhomogeneous perfect dielectric? Would it be possible to obtain the wave equation in that case?

**Q21.2.** Derive the Helmholtz equations, (21.7) and (21.8), from the wave equations, (21.5) and (21.6).

**Q21.3.** Write the expressions for at least three functions representing forward and backward traveling waves.

**Q21.4.** Is a plane electromagnetic wave with a component of the electric or magnetic field in the direction of propagation possible? Explain.

**Q21.5.** A perfect dielectric medium is not homogeneous, but $\epsilon$ is a smooth function of position, $\epsilon = \epsilon(x, y, z)$. Is a uniform plane wave possible in such a medium? Explain.

**Q21.6.** Write Eq. (21.10) in complex form and find its solutions.

**Q21.7.** What is the ratio of the wavelengths of a sinusoidal plane wave of frequency $f$ if it propagates in perfect dielectrics of permittivities $\epsilon_1$ and $\epsilon_2$, and permeability $\mu_0$?

**Q21.8.** What is the wavelength in a vacuum corresponding to the following frequencies of a plane wave: (1) 60 Hz, (2) 10 kHz, (3) 1 MHz, (4) 100 MHz, (5) 1 GHz, (6) 10 GHz, (7) 100 GHz, (8) 300 THz?

**Q21.9.** Does the expression $\mathbf{E} = E_1 \cos(\omega t - \beta z)\mathbf{u}_z$ represent a possible electric field of a plane wave? Explain.

**Q21.10.** Does the expression $\mathbf{E} = E_1 e^{j\beta x}\mathbf{u}_z$ represent a possible phasor expression for the electric field of a plane wave? Explain.

**Q21.11.** A circular loop of radius $a$ is situated in the field of a plane electromagnetic wave of wavelength $\lambda = a$. Is it possible in principle to evaluate the emf induced in the loop? If you think it is, can it be used for the evaluation of current intensity in the loop by means of circuit theory? Explain.

**Q21.12.** Does the concept of linear wave polarization make sense if the wave is not time-harmonic? What about the concept of circular and elliptical polarization? Explain.

**Q21.13.** A time-harmonic, linearly polarized plane wave propagates along the $z$ axis. Located along the $z$ axis is a row of small free charges. How do the charges move in time? Sketch their approximate position over a few time intervals.

**Q21.14.** Repeat question Q21.13 assuming that the polarization of the wave is circular.

**Q21.15.** What is the complex representation of the electric field of an elliptically polarized plane wave defined by Eqs. (21.30) and (21.31)?

**Q21.16.** Assuming that the wave propagates into the paper, is the polarization of the wave represented in the two sketches in Fig. Q21.16 right-handed or left-handed? Explain.
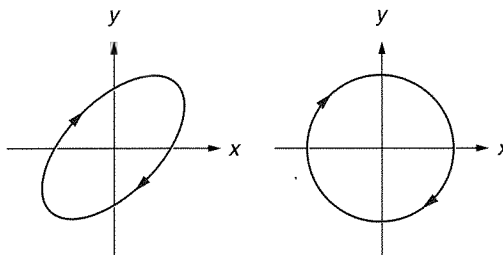


Figure Q21.16 Elliptic and circular polarization

**Q21.17.** Is the polarization of the wave represented in Fig. Q21.17 right-handed or left-handed? Explain.
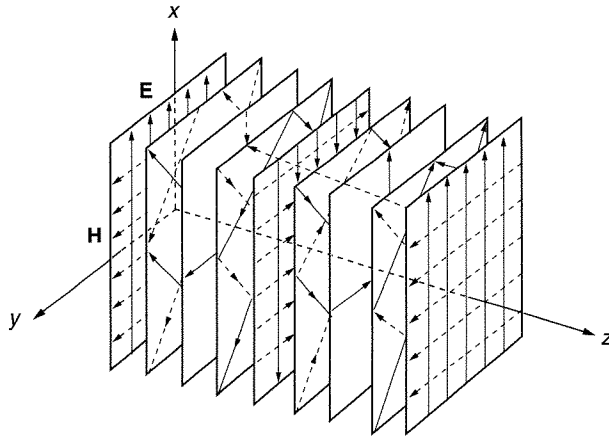


Figure Q21.17 Elliptically polarized wave

## PROBLEMS

**P21.1.** Assuming an $E_y$ component of the electric field, *derive* the corresponding component of the magnetic field in Eq. (21.20), following the same procedure as when $E_x$ was assumed.

**P21.2.** Prove that for a plane wave with both $E_x$ and $E_y$ components the vector **E** is normal to vector **H**. Evaluate the Poynting vector in that case.

**P21.3.** Repeat the entire derivation of the plane waves for time-harmonic plane waves and starting from complex forms of all the equations.

**P21.4.** A time-harmonic plane wave with an rms value of the electric field vector $E = 10\,\text{mV/m}$ propagates in a vacuum, and is normally incident on a screen that totally absorbs the energy of the wave. Find the absorbed energy per square meter of the screen in one hour.

**P21.5.** By measurements it was found that the time-average power of the sun's radiation on the surface of the earth is about $1.35\,\text{kW/m}^2$, for normal incidence of the plane waves from the sun. This radiation is composed of a very wide band of frequencies, and the components of different frequencies are generally polarized elliptically. Assuming, for simplicity, that the entire radiation is a linearly polarized wave of a single frequency, determine the rms value of its electric and magnetic field.

**P21.6.** The radius of the earth is about $6350\,\text{km}$. Assuming the entire energy of the sun's radiation reaching the earth is absorbed by the earth, calculate the time-average power of the absorbed energy, and the energy absorbed by the earth in one day. Compare this with the total man-produced energy, assuming that the time-average power of this energy during the day is about $12{,}500\,\text{GW}$.

**P21.7.** Due to various effects, human exposure to electromagnetic radiation is considered to be harmful above a certain time-average value of the Poynting vector. This estimated value depends on frequency, and differs greatly among different countries in the world. Assuming that above 10 GHz this value is on the order of $10\,\mathrm{mW/cm^2}$, compute the corresponding rms value of the electric and magnetic field of the plane wave with this time-average value of the Poynting vector. Compare this value of the electric field with the rms value of TV and broadcasting stations, which is on the order of mV/m.

**P21.8.** A circular wire loop of radius $a$ is situated in a vacuum in the electromagnetic field of a plane wave, of wavelength $\lambda$ ($\lambda \gg a$), and the rms value of the electric field strength $E$. How should the loop be positioned in order that the emf induced in it be maximal? Determine the rms value of the emf in that case.

**P21.9.** A rectangular wire loop with sides $a$ and $b$ is situated in a vacuum in the electromagnetic field of a time-harmonic plane wave. The amplitude of the electric field strength of the wave is $E$, and its wavelength is $\lambda$ ($\lambda \gg a, b$). The loop is oriented so that the maximal emf is induced in it. In case (1) the sides $a$ are parallel to the electric field of the wave, and in case (2) the sides $b$ are parallel to the electric field of the wave. Evaluate in both cases the emf (a) starting from Faraday's law of electromagnetic induction in its usual form ($e = -\mathrm{d}\Phi/\mathrm{d}t$), and (b) as an integral of the electric field strength of the wave around the contour.

**P21.10.** A sinusoidal plane wave, of frequency $f$ and time-average value of the Poynting vector $\mathcal{P}$, propagates through distilled water ($\mu = \mu_0$, $\epsilon = \epsilon_r\epsilon_0$). Find the rms value of the emf induced in a small circular loop of radius $a$, oriented so that the emf is maximal. What condition must be met in order that the loop can be considered as a quasi-static system?

**P21.11.** Prove that an elliptically polarized wave can be represented as a sum of two circularly polarized waves.

**P21.12.** Determine the time-average Poynting vector of a circularly polarized plane wave (1) starting from the expressions for the plane wave in time domain, and (2) starting from the phasor expressions for the plane wave.

**P21.13.** The electric field of a plane wave in complex (phasor) form is given by $\mathbf{E}(z) = (E_x\mathbf{u}_x + E_y\mathbf{u}_y)\mathrm{e}^{-\mathrm{j}\beta z}$. The components $E_x$ and $E_y$ are arbitrary complex numbers. Assuming that the wave propagates in the $+z$ direction, discuss the polarization of the wave, stating whether it is right-handed or left-handed for circular and elliptic polarization, if (1) $E_x = 1$, $E_y = 0$; (2) $E_x = 0$, $E_y = 5$; (3) $E_x = \mathrm{j}$, $E_y = -\mathrm{j}$; (4) $E_x = \mathrm{j}$, $E_y = 2$; (5) $E_x = (1 + \mathrm{j})$, $E_y = 0$; (6) $E_x = 1$, $E_y = \mathrm{j}$; or (7) $E_x = (1 + 2\mathrm{j})$, $E_y = (1 - \mathrm{j})$.

**P21.14.** Two plane waves of equal frequencies and phases propagate in the same $z$ direction. Both are circularly polarized, but in opposite directions (one is right-handed, the other left-handed). The amplitudes of the electric field strength of the two waves are $E_1$ and $E_2$. Find the polarization of the resultant wave in terms of $E_1$ and $E_2$, starting from the expressions of the waves in time domain.

**P21.15.** Two linearly polarized sinusoidal waves of the same frequency propagate in the $z$ direction. The electric field vectors of the two waves, $\mathbf{E}_1$ and $\mathbf{E}_2$, are along the $x$ and $y$ axes, respectively. Plot the trace that the tip of the resulting vector, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$, traces at $z = 0$ in time, as a function of the ratio of amplitudes $E_1$ and $E_2$, and their relative phase, $\phi$.

# 22

# *Reflection and Refraction of Plane Waves*

## 22.1 Introduction

In reality, plane electromagnetic waves frequently encounter obstacles along their propagation paths: hills, buildings, metallic antennas aimed at receiving the messages the waves carry, objects from which they are supposed to partly reflect (as when the wave is a radar beam), and so on. In such cases, the wave induces conduction currents in the object (if the object is metallic), or polarization currents (if the object is made of an insulator). These currents are, of course, sources of a secondary electromagnetic field. This field is known as the *scattered field*, and the process that creates it is known as *scattering of electromagnetic waves*. The objects, or obstacles, are called *scatterers*.

The determination of scattered fields is a difficult problem even in the case of simple scatterers, and can rarely be solved analytically. Numerical analysis offers various solutions. There is one class of problems, however, for which the determination of the scattered field is remarkably simple. When a *plane* electromagnetic wave is incident on a *planar* boundary between two *homogeneous* media, the scattered waves are also plane waves. One of these waves is radiated back into the half-space of the incident wave: this wave is known as the *reflected wave*. There is also a wave in the other half-space (except in the case of a perfect conductor), propagating generally

411

in a different direction from the incident wave; it is therefore called the *refracted or transmitted wave*.

Naturally, the described geometry is an idealized one. Nevertheless, the results we will arrive at have great practical importance because many real problems can be solved in this manner with sufficient accuracy.

## 22.2 Plane Waves Normally Incident on a Perfectly Conducting Plane

The simplest case of wave reflection is when a uniform plane wave is incident normally on the planar interface between a perfect dielectric, of parameters $\epsilon$ and $\mu$, and a perfect conductor. Let the interface be at $z = 0$, and let the wave of angular frequency $\omega$ have $E_x$ and $H_y$ components, as indicated in Fig. 22.1. We wish to determine the resulting wave for $z \leq 0$. (We know that inside the perfect conductor there is no field.)

The physics of wave scattering in this case is fairly obvious. The incident wave induces currents and charges only on the surface of the perfect conductor. (For a perfect conductor, the skin depth is infinitely small.) Since inside the conductor there is no field, we can consider this layer of currents and charges to exist in a homogeneous dielectric of parameters $\epsilon$ and $\mu$. The distribution of these sources must be such that their field exactly cancels the incident field inside the conductor (that is, for $z > 0$). So we know that the scattered field for $z > 0$ is exactly the same in amplitude as the incident field, but is $\pi$ out of phase. The current sheet obviously produces a symmetrical field in the half-space $z < 0$, that is, a plane wave propagating back in the $-z$ direction. This reradiated wave is the reflected wave. From this reasoning, the reflected wave has the same amplitude as the incident wave. At $z = 0$, its $E$-field vector is the same as that of the incident field, but in the opposite direction.

To put these conclusions into equations, let the incident wave (in phasor form) be represented by

$$\mathbf{E}_i(z) = E e^{-j\beta z}\mathbf{u}_x \qquad \mathbf{H}_i(z) = H e^{-j\beta z}\mathbf{u}_y, \tag{22.1}$$
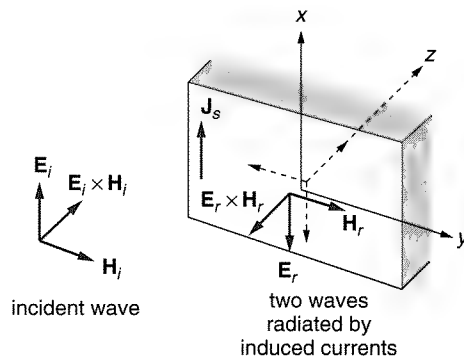


Figure 22.1 Uniform plane wave normally incident on the planar interface between a perfect dielectric and perfect conductor

where $E/H = \eta$ ($\eta = \sqrt{\mu/\epsilon}$, the intrinsic impedance of the medium). The reflected wave is then of the form

$$\mathbf{E}_r(z) = -Ee^{+j\beta z}\mathbf{u}_x \qquad \mathbf{H}_r(z) = He^{+j\beta z}\mathbf{u}_y. \tag{22.2}$$

The Poynting vector (which represents power flow) for the reflected wave is $-z$ oriented, which determines the sign of $\mathbf{H}_r(z)$.

The total field for $z < 0$ is obtained as a superposition of the waves in Eqs. (22.1) and (22.2):

$$\mathbf{E}_{\text{tot}}(z) = \mathbf{E}_i(z) + \mathbf{E}_r(z) = E\left(e^{-j\beta z} - e^{+j\beta z}\right)\mathbf{u}_x = -2jE\sin\beta z\,\mathbf{u}_x, \tag{22.3}$$

$$\mathbf{H}_{\text{tot}}(z) = \mathbf{H}_i(z) + \mathbf{H}_r(z) = H\left(e^{-j\beta z} + e^{+j\beta z}\right)\mathbf{u}_y = 2H\cos\beta z\,\mathbf{u}_y. \tag{22.4}$$

The instantaneous values of the two vectors are therefore

$$\mathbf{E}_{\text{tot}}(z,t) = 2E\sqrt{2}\sin\beta z\cos(\omega t - \pi/2)\,\mathbf{u}_x = 2E\sqrt{2}\sin\beta z\sin\omega t\,\mathbf{u}_x, \tag{22.5}$$

$$\mathbf{H}_{\text{tot}}(z,t) = 2H\sqrt{2}\cos\beta z\cos\omega t\,\mathbf{u}_y. \tag{22.6}$$

The total wave does *not* contain the factor $e^{\pm j\beta z}$. We already had such a case for voltage and current along open and shorted transmission lines. We know that it is not a progressive, traveling wave in either direction, but a *standing wave*. As along such transmission lines, there are planes in which $\mathbf{E}_{\text{tot}}(z,t)$ is zero at all times. These planes are defined by $\beta z = -n\pi$, $n = 0, 1, 2, \ldots$. Similarly, the magnetic field is zero at all times in planes defined by $\beta z = -(2n+1)\pi/2$, $n = 0, 1, 2, \ldots$. Thus, the total wave actually stays where it is, only pulsating in time according to the sine law (the $E$ field) or the cosine law (the $H$ field). The expressions describing a wave in which the time and space coordinates are as in Eqs. (22.3) and (22.4) in complex notation, or as in Eqs. (22.5) and (22.6) in the time domain, always represent standing waves. A sketch of instantaneous values of the total $E$ and $H$ field in front of the interface, for $\omega t = 0, \pi/4, 2\pi/4$, and $3\pi/4$, is shown in Fig. 22.2.

**Example 22.1—The Fabry-Perot resonator.** Consider again the case of a plane wave reflecting off a perfectly conducting plane, which behaves like a mirror. Note that according to Eq. (22.5), $\mathbf{E}_{\text{tot}}(-n\lambda/2, t) = 0$ *at all times* in the planes $z = -n\lambda/2$, $n = 1, 2, \ldots$. The electric field vector is tangential to these planes. Therefore, if we insert a perfectly conducting sheet (also a mirror) in any of these planes (i.e., for any $n$), nothing will change, since the boundary condition on the plane for vector $\mathbf{E}$ is satisfied automatically. In this manner, we obtained a semi-infinite region to the left of the sheet with the standing wave, and a region between the original mirror and the sheet in which *the electric and magnetic fields oscillate* as in Fig. 22.2. When the electric field is maximum, the magnetic field is zero, and conversely. This means that in the region between the two mirrors, the electric energy is being converted into magnetic energy, and vice versa. This is typical behavior for resonant electric circuits. We can conclude that from the energy point of view, just like in an *LC* circuit, this is a resonator, but a *spatial resonator*. This particular type of spatial resonator is known as the *Fabry-Perot resonator*, and is used extensively in optics and at millimeter-wave and infrared frequencies.

The Fabry-Perot resonator has a very useful property. Note that losses exist only in the original plane and in the sheet, due to the skin effect and finite conductivity of the metal. The
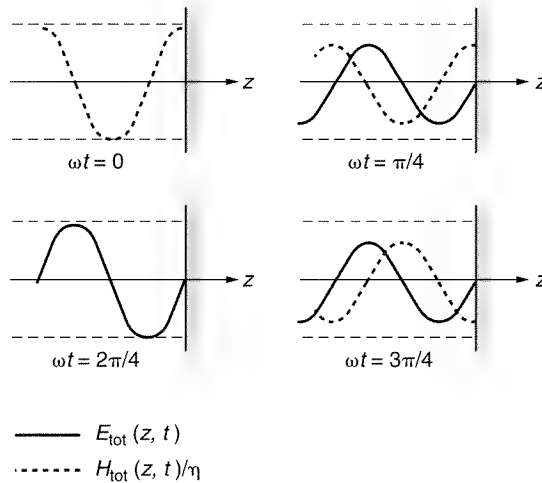
$$\text{——— } E_{tot}\,(z,\,t)$$
$$\text{······ } H_{tot}\,(z,\,t)/\eta$$

**Figure 22.2** Sketch of instantaneous values of $E_{tot}(z,t)$ and $H_{tot}(z,t)$ in Eqs. (22.5) and (22.6) for a standing wave

electromagnetic energy located in the resonator, however, increases with the number $n$, that is, with the number of half wavelengths of the wave contained in the region between the two mirrors. At high frequencies (e.g., in the microwave region or in optics), $n$ can be made very large with reasonable dimensions of the resonator. (For example, at 30 GHz, the wavelength is 1 cm, and a 10-cm-long resonator has $n = 20$.) Therefore, the Fabry-Perot resonator can have an arbitrarily large ratio between the energy stored in the resonator and losses in one cycle. We know that this ratio is proportional to the quality factor of the resonator (the $Q$ factor). Therefore, the $Q$ factor of a Fabry-Perot resonator can be extremely large (on the order of tens of thousands) when compared with that of a resonant circuit (which has a maximum $Q$ factor of about one hundred). Of course, due to the finite size of the plates in reality, there will always be some leakage of electromagnetic energy from the resonator, which we do not take into account in this simplified analysis. Also, usually energy is purposely taken out of the resonator: for example, one of the mirrors may be partially transparent so that not all of the energy is reflected back into the cavity. This also is not taken into account in our simplified analysis.

*Questions and problems:*    Q22.1 to Q22.4, P22.1 to P22.6

## 22.3 Reflection and Transmission of Plane Waves Normally Incident on a Planar Boundary Surface Between Two Dielectric Media

Let us consider two lossless dielectric media, 1 and 2, of parameters $\epsilon_1$ and $\mu_1$, and $\epsilon_2$ and $\mu_2$, respectively, separated by a planar interface, as in Fig. 22.3. Let the incident wave, with an electric field $E_{1i}$ and of angular frequency $\omega$, propagate in medium 1 toward the interface, normal to it, with the vector $\mathbf{E}$ parallel to the $x$ axis (Fig. 22.3).
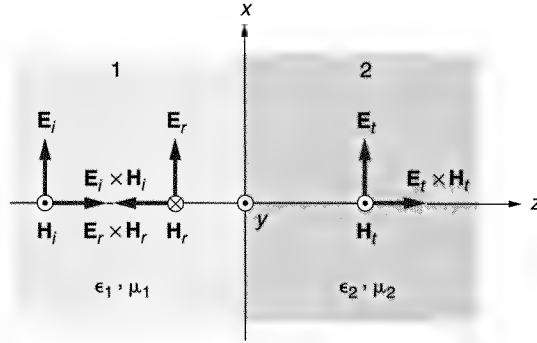
Figure 22.3 Two lossless dielectric media divided by a plane interface. The incident wave from medium 1 is partly reflected, and partly transmitted to medium 2.

A part of the incident electromagnetic energy will be reflected from the interface, and a part will be transmitted into medium 2. Assume the reference directions of the $E$ field for the reflected and transmitted waves as indicated. The reference directions of the $H$ field for the three waves are then as shown in the figure.

We wish to determine the relative intensities $E_{1r}$ and $E_2$ at $z = 0$ of the $E$ field for the reflected and transmitted waves. For this, we first need the expressions for these fields. With the adopted reference directions of the vectors in Fig. 22.3, they have the forms

$$\mathbf{E}_i(z) = E_{1i}e^{-j\beta_1 z}\mathbf{u}_x, \quad \mathbf{H}_i(z) = \frac{E_{1i}}{\eta_1}e^{-j\beta_1 z}\mathbf{u}_y, \tag{22.7}$$

$$\mathbf{E}_r(z) = E_{1r}e^{+j\beta_1 z}\mathbf{u}_x, \quad \mathbf{H}_r(z) = -\frac{E_{1r}}{\eta_1}e^{+j\beta_1 z}\mathbf{u}_y, \tag{22.8}$$

$$\mathbf{E}_t(z) = E_2 e^{-j\beta_2 z}\mathbf{u}_x, \quad \mathbf{H}_t(z) = \frac{E_2}{\eta_2}e^{-j\beta_2 z}\mathbf{u}_y. \tag{22.9}$$

We can now write the boundary conditions, i.e., the requirements that the tangential components of the total vectors $\mathbf{E}$ and $\mathbf{H}$ on two sides of the interface be the same:

$$E_{1i} + E_{1r} = E_2, \quad \frac{E_{1i}}{\eta_1} - \frac{E_{1r}}{\eta_1} = \frac{E_2}{\eta_2}. \tag{22.10}$$

By solving these two equations for $E_{1r}$ and $E_2$, we obtain

$$E_{1r} = \frac{\eta_2 - \eta_1}{\eta_1 + \eta_2}E_{1i}, \quad E_2 = \frac{2\eta_2}{\eta_1 + \eta_2}E_{1i}. \tag{22.11}$$

Note that these are the same expressions we found in Chapter 18 for the incident (forward) and reflected (backward) voltages along a line of characteristic impedance $Z_1$ terminated in an infinite line of characteristic impedance $Z_2$ (Example 18.8). As

in the case of transmission lines, the ratio $E_{1r}/E_{1i}$ is known as the *reflection coefficient*, and the ratio $E_2/E_{1i}$ the *transmission coefficient*:

$$\rho = \frac{\eta_2 - \eta_1}{\eta_1 + \eta_2} \qquad \text{(the reflection coefficient, dimensionless)} \tag{22.12}$$

$$\tau = \frac{2\eta_2}{\eta_1 + \eta_2} \qquad \text{(the transmission coefficient, dimensionless).} \tag{22.13}$$

Note also that $\rho$ and $\tau$ as just derived are defined with respect to the same reference directions of all three components of the *electric field* of the three waves.

In medium 2 there is only the progressive transmitted wave. In medium 1, however, we have the incident wave and the reflected wave, the total field being the sum of the two:

$$\mathbf{E}_1(z) = \mathbf{E}_i(z) + \mathbf{E}_r(z) = E_{1i}e^{-j\beta_1 z}\left(1 + \rho e^{+j2\beta_1 z}\right)\mathbf{u}_x. \tag{22.14}$$

This is the same expression as for the voltage along a transmission line terminated in a load, Eq. (18.22a). The electric field in medium 1 is therefore of the same form as the voltage in the analogous transmission-line case. The following analysis, which parallels that from Chapter 18, shows this clearly.

If $\rho > 0$ (that is, if $\eta_2 > \eta_1$), the expression in parentheses is the largest, equal to $(1 + \rho)$, in planes defined by the following equation (note that medium 1 occupies the half-space $z < 0$):

$$2\beta_1 z_{\max} = -2n\pi, \qquad \text{or} \qquad z_{\max} = -\frac{n\pi}{\beta_1} = -\frac{n\lambda_1}{2} \qquad n = 0, 1, \ldots. \tag{22.15}$$

This expression is minimal, equal to $(1 - \rho)$, in planes

$$z_{\min} = -(2n + 1)\frac{\lambda_1}{4} \qquad n = 0, 1, \ldots. \tag{22.16}$$

If $\rho < 0$ (that is, if $\eta_2 < \eta_1$), $z_{\max}$ and $z_{\min}$ simply exchange places.

The resultant wave in medium 1 can be visualized as a sum of a progressive wave of rms value $(1 - |\rho|)E_{1i}$ and a standing wave of rms value (in the maximum of the standing wave) $2|\rho|E_{1i}$. This becomes evident if Eq. (22.14) is rewritten in the form

$$\mathbf{E}_1(z) = (1 - \rho)E_{1i}e^{-j\beta_1 z}\mathbf{u}_x + 2\rho E_{1i}\cos\beta_1 z\,\mathbf{u}_x \qquad (\rho > 0), \tag{22.17}$$

namely,

$$\mathbf{E}_1(z) = (1 + \rho)E_{1i}e^{-j\beta_1 z}\mathbf{u}_x + 2j\rho E_{1i}\sin\beta_1 z\,\mathbf{u}_x \qquad (\rho < 0). \tag{22.18}$$

The ratio analogous to the voltage standing-wave ratio, VSWR, from transmission lines,

$$\text{SWR} = \frac{|E_1(z)|_{\max}}{|E_1(z)|_{\min}} = \frac{1 + |\rho|}{1 - |\rho|} \qquad \text{(standing-wave ratio, dimensionless)} \tag{22.19}$$

is known as the *standing-wave ratio*. Since $|\rho| < 1$, it increases with the increase of $|\rho|$.

*Questions and problems:*   Q22.5, P22.7 to P22.11
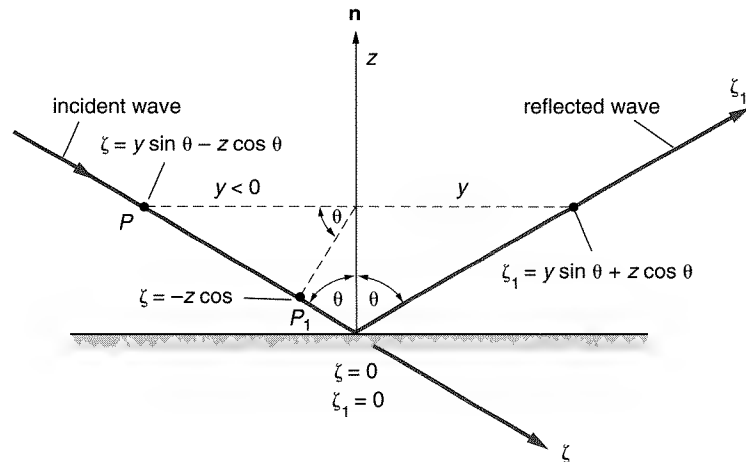
Figure 22.4 The direction of propagation of a wave reflected off a perfectly conducting plane is at the same angle with respect to the normal to the plane as the incident wave direction.

## 22.4 Plane Waves Obliquely Incident on a Perfectly Conducting Plane

Assume that a uniform plane wave propagating in a perfect dielectric of parameters $\epsilon$ and $\mu$ is obliquely incident on a perfectly conducting plane. As in the case of normal incidence, the scattered field due to the currents and charges induced on the plane must be such that it cancels the incident field in the perfect conductor. Thus, these currents and charges will produce a wave inside the conductor. This wave will be exactly the same as the incident wave, propagating in the same direction, but of opposite phase. The same field will be produced on the other side of the plane, resulting in a "reflected wave." Therefore, the direction of propagation of the reflected wave will be at the same angle $\theta$ (with respect to the normal to the plane) as the incident wave (Fig. 22.4).

The plane containing the vector $\mathbf{n}$ and the directions of propagation of the incident and reflected waves is known as the *plane of incidence*. Any incident plane wave can be represented as a superposition of two plane waves, one with the vector $\mathbf{E}$ normal to the plane of incidence, and the other with the vector $\mathbf{E}$ parallel to it. These two cases are simpler to analyze than any other. Therefore, we consider these two special cases only, knowing that any other case can be obtained by superposition.

### 22.4.1  VECTOR E NORMAL TO THE PLANE OF INCIDENCE

It is customary to say that this wave has *normal* or *horizontal polarization*. (The term "normal" refers to the plane of incidence, and "horizontal" refers to the fact that frequently the reflection plane is the earth's surface, in which case the vector $\mathbf{E}$ of this wave is horizontal.) The case of a horizontally polarized incident wave is sketched in Fig. 22.5, with the adopted reference directions for vectors $\mathbf{E}$ and $\mathbf{H}$.
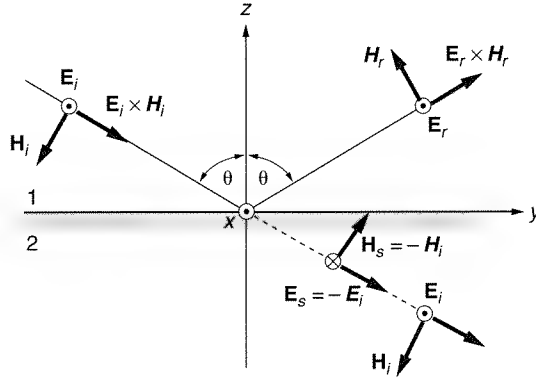
**Figure 22.5** A horizontally polarized plane wave
reflected from a perfectly conducting plane

The wave propagates along an axis ($\zeta$ axis, Fig. 22.4) not coinciding with a coordinate axis $x$, $y$, or $z$. To write the expression for the wave in terms of the rectangular coordinates, we need to determine the distance of a point $P$ (Fig. 22.4) from the origin of the $\zeta$ axis ($\zeta = 0$) in terms of $x$, $y$, and $z$. From Fig. 22.4 it is seen that for the incident wave this distance equals $\zeta = y \sin\theta - z \cos\theta$. For $P$ note that $y < 0$ and $z > 0$, and that $P$ is in the negative part of the $\zeta$ axis. So the factor $e^{-j\beta\zeta}$ (the factor for the wave propagating in the direction of the $\zeta$ axis) becomes $e^{-j\beta(y\sin\theta - z\cos\theta)}$. The expression for the complex electric field of the incident wave is thus

$$\mathbf{E}_i(y, z) = E e^{-j\beta(y\sin\theta - z\cos\theta)} \mathbf{u}_x. \tag{22.20}$$

In the same wave we conclude that the $E$ field of the reflected wave is given by

$$\mathbf{E}_r(y, z) = -E e^{-j\beta(y\sin\theta + z\cos\theta)} \mathbf{u}_x. \tag{22.21}$$

The minus sign comes from the requirement that the total tangential $E$ field on the plane $z = 0$ must be zero for any $y$.

The total electric field has only an $x$ component, given by

$$E_{\text{tot}}(y, z) = E_i(y, z) + E_r(y, z) = E e^{-j\beta y\sin\theta} \left( e^{j\beta z\cos\theta} - e^{-j\beta z\cos\theta} \right),$$

from which

$$E_{\text{tot}}(y, z) = 2jE \sin(\beta z \cos\theta) e^{-j\beta y\sin\theta}. \tag{22.22}$$

We see that the total electric field is a standing wave in the $z$ direction, and a traveling wave in the $y$ direction. The wavelength in the $z$ direction is given by

$$\lambda_z = \frac{2\pi}{\beta\cos\theta} = \frac{\lambda}{\cos\theta}, \tag{22.23}$$

where $\lambda$ is the wavelength of the incident (and reflected) wave. The vector $\mathbf{E}$ is zero in the planes in which $\beta z \cos\theta = n\pi$, $n = 0, 1, 2, \ldots$, or

$$z_{E=0} = \frac{n\lambda_z}{2} = \frac{n\lambda}{2\cos\theta}, \quad n = 0, 1, 2, \ldots. \tag{22.24}$$

In the direction of the $y$ axis, the total field behaves as a traveling wave, with a phase velocity along the $y$ axis

$$v_{\text{ph}} = \frac{\omega}{\beta_y} = \frac{\omega}{\beta \sin\theta} = \frac{c}{\sin\theta}, \qquad c = \frac{1}{\sqrt{\epsilon\mu}} \tag{22.25}$$

(note that the phase coefficient with respect to the $y$ axis is the entire factor of j$y$ in the exponent, that is, $\beta_y = \beta \sin\theta$), and with a wavelength along the $y$ axis

$$\lambda_y = \frac{2\pi}{\beta_y} = \frac{\lambda}{\sin\theta}. \tag{22.26}$$

**Example 22.2—The rectangular waveguide.** Because in the planes $z_{E=0}$ the magnitude of the $E$ field is zero at all times, we can insert into any one of these planes a perfectly conducting sheet. Assume that in some way we switch the field above the sheet off. What remains is a system of two perfectly conducting planes guiding a specific wave propagating in the $y$ direction.

We can go a step further. The vector **E** is normal to the planes defined by $x =$ constant. Introducing a perfectly conducting sheet in one or more such planes will not change the field— it will only induce surface charges of opposite signs on the two faces of the sheet. However, if we imagine two such sheets together with the first sheet parallel to the plane $z = 0$, we obtain a rectangular tube through which an electromagnetic wave propagates just like water flows through a pipe. Such a rectangular metallic tube is known as the *rectangular waveguide*. It is used extensively for guiding electromagnetic energy at microwave and millimeter-wave frequencies.

We will learn in the next chapter that this type of electromagnetic wave is only one of an infinite number of wave types that can propagate through such rectangular metallic tubes.

**Example 22.3—Determination of the total $H$ field.** With reference to Fig. 22.5, the total $H$ field has two components, $H_y$ and $H_z$. Let us determine them as an exercise. The two components of the total $H$ field are obtained as the sum of these components for the incident and the reflected waves:

$$H_{\text{tot } y}(y, z) = H_{iy}(y, z) + H_{ry}(y, z)$$

$$= -\frac{E}{\eta}e^{-j\beta(y\sin\theta - z\cos\theta)}\cos\theta - \frac{E}{\eta}e^{-j\beta(y\sin\theta + z\cos\theta)}\cos\theta.$$

After simple rearrangements similar to those in deriving Eq. (22.22), we obtain

$$H_{\text{tot } y}(y, z) = -2\frac{E}{\eta}\cos\theta\cos(\beta z\cos\theta)e^{-j\beta y\sin\theta}.$$

The $H_z$ component is obtained in a similar way, which is left as an exercise for the reader. The result is

$$H_{\text{tot } z}(y, z) = 2j\frac{E}{\eta}\sin\theta\sin(\beta z\cos\theta)e^{-j\beta y\sin\theta}.$$

Note that $H_z$ is zero on the perfectly conducting plane, as it should be (the magnetic field can have no normal component on a perfect conductor—see Example 20.2).

### 22.4.2   VECTOR E PARALLEL TO THE PLANE OF INCIDENCE

Assume now that vector **E** is parallel to the plane of incidence, as sketched in Fig. 22.6. Because the tangential component ($y$ component) at the plane must be zero, again the reflected wave has the same amplitude. The directions of the $E$ and $H$ vectors indicated in the figure represent their reference directions.

We now have two $E$-field components of the incident and the reflected waves, the $y$ and the $z$ components. Both must be of the form in Eqs. (22.20) and (22.21):

$$E_{iy}(y, z) = E \cos \theta e^{-j\beta(y \sin \theta - z \cos \theta)}, \tag{22.27}$$

$$E_{iz}(y, z) = E \sin \theta e^{-j\beta(y \sin \theta - z \cos \theta)}, \tag{22.28}$$

and

$$E_{ry}(y, z) = -E \cos \theta e^{-j\beta(y \sin \theta + z \cos \theta)}, \tag{22.29}$$

$$E_{rz}(y, z) = E \sin \theta e^{-j\beta(y \sin \theta + z \cos \theta)}. \tag{22.30}$$

The total components are the sum of these:

$$E_{\text{tot }y}(y, z) = E_{iy}(y, z) + E_{ry}(y, z) = 2jE \cos \theta \sin(\beta z \cos \theta)e^{-j\beta y \sin \theta}, \tag{22.31}$$

$$E_{\text{tot }z}(y, z) = E_{iz}(y, z) + E_{rz}(y, z) = 2E \sin \theta \cos(\beta z \cos \theta)e^{-j\beta y \sin \theta}. \tag{22.32}$$

**Example 22.4—Determination of the total H field.** With reference to Fig. 22.6, the total $H$ field in this case has the $x$ component only. Because we know that $H = E/\eta$, the expressions
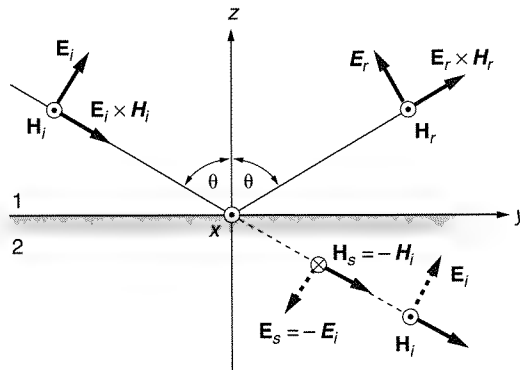


Figure 22.6 Reflection of a vertically polarized plane wave from a perfectly conducting plane

for the $H$ field of the incident and reflected waves are

$$H_{ix}(y, z) = \frac{E}{\eta} e^{-j\beta(y\sin\theta - z\cos\theta)} \qquad \text{and} \qquad H_{rx}(y, z) = \frac{E}{\eta} e^{-j\beta(y\sin\theta + z\cos\theta)}.$$

The total $H$ field is hence

$$H_{\text{tot }x}(y, z) = 2\frac{E}{\eta} \cos(\beta z \cos\theta) e^{-j\beta y \sin\theta}.$$

**Example 22.5—Maximal emf induced in a small loop above a perfectly conducting plane.** Assume that we wish to receive a signal contained in the incident wave. One way, which is quite easy to understand, is to use a loop of wire (e.g., a circular one of radius $a$) much smaller than the wavelength of the wave. The emf induced in the loop is then obtained according to Faraday's law. So to obtain a maximal emf, the principal thing we have to determine is where the magnetic field is maximal, and what its direction is at that point.

The magnetic field has a maximum at $z = 0$, with a value equal to

$$H_{\text{tot }x}(y, 0) = 2\frac{E}{\eta} e^{-j\beta y \sin\theta}.$$

The last factor in this expression determines just the initial phase of the field along the $y$ axis. To simplify, let $y = 0$. The maximal possible complex emf [note that the complex counterpart of the expression $e(t) = -d\Phi(t)/dt$ is $\underline{e} = -j\omega\underline{\Phi}$] induced in the loop is thus

$$\mathcal{E}_{\max} = -2j\frac{E}{\eta}\omega\mu_0 a^2\pi.$$

In this case, we have used the small loop as a receiving antenna. This type of antenna, which develops a voltage between its terminals due to a time-variable flux of the magnetic field through its contour, is called a *loop antenna.* We could also have used two short straight wires connected to a voltmeter that measures the emf. In this case, there is an induced emf due to the integral of the induced electric field along the wires. This type of antenna is called a *short dipole antenna.*

*Questions and problems:*   Q22.6 and Q22.7, P22.12 and P22.13

## 22.5   Reflection and Transmission of Plane Waves Obliquely Incident on a Planar Boundary Surface Between Two Dielectric Media

When a plane wave is obliquely (at an angle) incident on a plane interface between two media, the formulation of boundary conditions becomes more complex than when incidence is normal. Of course, again a part of the incident energy is reflected back into medium 1 (of parameters $\epsilon_1$ and $\mu_1$), and a part is transmitted into medium 2 (of parameters $\epsilon_2$ and $\mu_2$). We shall see that the direction of propagation of the reflected wave makes the same angle with the normal to the interface as the incident wave, as before. However, the transmitted wave is deflected with respect to this normal. The transmitted wave in this case is therefore frequently termed the *refracted wave.*

The amplitudes of the reflected and refracted waves depend, among other things, on the polarization of the wave (i.e., on the electric field vector being parallel or normal to the plane of incidence). However, the angles that the direction of propagation of the two secondary waves make with the normal to the interface are the same for any polarization.

Figure 22.7 shows equiphase planes (planes of constant phase) and the directions of propagation of the incident, reflected, and refracted waves. These planes in medium 1 are moving with a velocity $c_1 = 1/\sqrt{\epsilon_1\mu_1}$, and in medium 2 with a velocity $c_2 = 1/\sqrt{\epsilon_2\mu_2}$. Indicated in the figure are a few equiphase planes of the three waves. Let the boundary conditions be satisfied at the instant for which Fig. 22.7 is valid. In order that they remain satisfied at all times, it is necessary that the relative amplitudes and phases of the three waves at the interface remain unchanged. This is possible only if the intersections of the equiphase planes with the interface move *along the interface* at the same speed.

From Fig. 22.7, this velocity for the incident and reflected wave is $c_1/\sin\theta_i$ and $c_1/\sin\theta_r$, and for the refracted wave $c_2/\sin\theta_2$. To satisfy this condition we conclude that, first, $\theta_r = \theta_i$. This angle we shall therefore denote as $\theta_1$. Second, the condition $c_1/\sin\theta_1 = c_2/\sin\theta_2$ must also hold, so

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{c_1}{c_2} = \sqrt{\frac{\epsilon_2\mu_2}{\epsilon_1\mu_1}}. \tag{22.33}$$

*(Snell's law)*
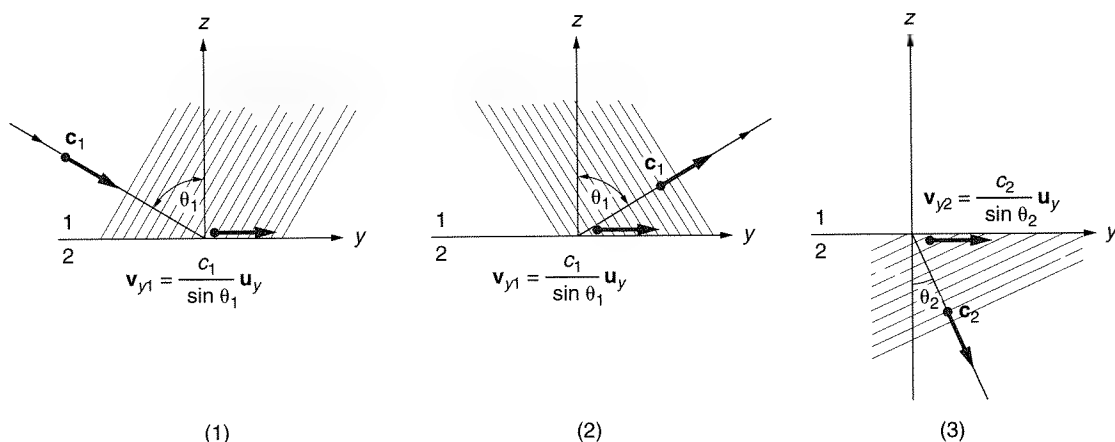


Figure 22.7 Equiphase planes and the directions of propagation of the incident, reflected, and refracted waves

This relation is known as *Snell's law*. The ratio $c_1/c_2$ is termed the *index of refraction* for media 1 and 2, and is often denoted as $n_{12}$, especially in optics. If $\mu_1 = \mu_2 = \mu_0$ (which is most often the case),

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{c_1}{c_2} = \sqrt{\frac{\epsilon_2}{\epsilon_1}} \qquad (\mu_1 = \mu_2). \tag{22.34}$$

We know that if $0 \leq \beta < \alpha \leq \pi/2$, then $\sin \alpha > \sin \beta$. Snell's law and Eq. (22.34) therefore tell us that for $\epsilon_1 < \epsilon_2$, $\theta_1 > \theta_2$. This means that the wave is refracted toward the normal. The refracted wave exists for any $\theta_1$.

If $\epsilon_1 > \epsilon_2$, however, $\theta_1 < \theta_2$. This means that the direction of propagation of the refracted wave makes a greater angle with the normal than that of the incident wave. So for a certain angle $\theta_1$ the angle $\theta_2$ will become $\pi/2$, and cannot increase further. From Eq. (22.34), this limiting angle $\theta_1 = \theta_t$ is defined by

$$\frac{\sin \theta_t}{\sin(\pi/2)} = \sin \theta_t = \sqrt{\frac{\epsilon_2}{\epsilon_1}} \qquad (\epsilon_1 > \epsilon_2, \mu_1 = \mu_2). \tag{22.35}$$

*[Sine of critical angle (angle of total reflection)]*

This particular angle $\theta_1 = \theta_t$ is known as the *critical angle*, or the *angle of total reflection*.

For $\theta_1 > \theta_t$, the sine of $\theta_2$ must be *greater than one* in order for the boundary conditions to be satisfied. At first glance it might seem as if we made a mistake. The sine of a *real* angle cannot be greater than one. However, the sine of a complex angle can be larger than unity. This is easily understood if we set $\theta_2 = \pi/2 - jx$ and recall the expression for the sine in terms of the exponential function:

$$\sin(\pi/2 - jx) = \frac{1}{2j} \left[ e^{j(\pi/2 - jx)} - e^{-j(\pi/2 - jx)} \right] = \frac{1}{2j} \left( je^x + je^{-x} \right) = \cosh x,$$

since $e^{\pm j\pi/2} = \pm j$. The *hyperbolic consine* function of $x$, $\cosh x = (e^x + e^{-x})/2$ can have any positive value between one and infinity.

What happens then if $\theta_1 > \theta_t$? Obviously, there can be no refracted wave in medium 2, so all of the energy of the incident wave is reflected back into medium 1. Example 22.9 will show that indeed, the magnitude of the reflection coefficient is then equal to one. This is known as *total reflection*. It has many applications and is encountered on many occasions.

**Example 22.6—Apparent shape of an oar observed from a rowboat.** If you are in a rowboat on clear, calm water, and observe the oar immersed in the water, the oar looks as if it is broken at the water level: the immersed part of the oar appears higher than expected. This is easy to explain using Snell's law. You see the immersed part of the oar because light rays, i.e., electromagnetic waves, are reflected from the oar toward your eyes. They pass the water-air interface and are refracted in the air away from the normal because the permittivity of water is greater than that of air. Therefore, the oar looks broken.

If you observe the oar from a distant point, you will not be able to see the submerged part of the oar. This is because the rays from the submerged part of the oar in that case are incident on the water-air interface at an angle greater than the critical angle, and there are no transmitted rays in the air in your direction anymore.

*Questions and problems:*    Q22.8, P22.14

## 22.6    Fresnel Coefficients

Snell's law and the phenomenon of total reflection are valid for any polarization of the incident wave. The reflection and transmission coefficients, which are defined in the same way as for normal incidence, are different for normal and parallel polarization. In this section we derive the so-called *Fresnel coefficients*, which are reflection and transmission coefficients written in terms of the angle of incidence and the material properties (wave impedances) of the two media.

For waves obliquely incident on the interface between two dielectrics, we need to consider the two polarizations separately, similarly to the conductor case.

### 22.6.1    VECTOR E NORMAL TO THE PLANE OF INCIDENCE (TRANSVERSE ELECTRIC CASE)

Let the reference directions of the field vectors of the incident, reflected, and refracted waves be as in Fig. 22.8. Let $E_{1i}$, $E_{1r}$, and $E_2$ and $H_{1i}$, $H_{1r}$, and $H_2$ be the complex rms values of the vectors of the three waves at the interface ($z = 0$). The boundary conditions require that the tangential components of the total $E$ field and the total
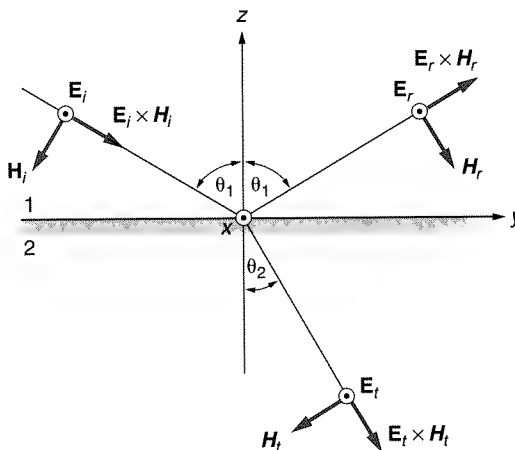


Figure 22.8 Reference directions of the field vectors of the incident, reflected, and refracted waves for a normally polarized incident wave

$H$ field on the two sides of the interface be equal. This results in the following two equations:

$$E_{1i} + E_{1r} = E_2 \qquad (H_{1i} - H_{1r}) \cos \theta_1 = H_2 \cos \theta_2. \qquad (22.36)$$

Since $H_{1i} = E_{1i}/\eta_1$, $H_{1r} = E_{1r}/\eta_1$, and $H_2 = E_2/\eta_2$, we have two linear equations in two unknowns, $E_{1r}$ and $E_2$. Solving these equations we obtain

$$\rho_n = \left( \frac{E_{1r}}{E_{1i}} \right)_n = \frac{\eta_2 \cos \theta_1 - \eta_1 \cos \theta_2}{\eta_2 \cos \theta_1 + \eta_1 \cos \theta_2}, \qquad (22.37)$$

$$\tau_n = \left( \frac{E_2}{E_{1i}} \right)_n = \frac{2\eta_2 \cos \theta_1}{\eta_2 \cos \theta_1 + \eta_1 \cos \theta_2}. \qquad (22.38)$$

*[Fresnel's coefficients for normal (TE) polarization]*

The reflection and transmission coefficients, $\rho_n$ and $\tau_n$, are known as the *Fresnel coefficients* for normal polarization. They are also sometimes called the *transverse electric (TE)* Fresnel coefficients. In these expressions, according to Snell's law in Eq. (22.33), $\cos \theta_2$ must be calculated as

$$\cos \theta_2 = \sqrt{1 - \sin^2 \theta_2} = \frac{c_2}{c_1} \sqrt{\left( \frac{c_1}{c_2} \right)^2 - \sin^2 \theta_1}. \qquad (22.39)$$

The expressions for the $\rho$ and $\tau$ coefficients are general. For perfect dielectrics, having real intrinsic impedances, they are real. As a consequence, the reflected wave on the interface is either in phase (if $\rho > 0$) or in counterphase (if $\rho < 0$) with respect to the incident wave. If either of the two media is not a perfect dielectric, the intrinsic impedance of that medium is complex, so that both $\rho$ and $\tau$ are complex as well, and the phase difference between the field vectors on the interface is different from $\pi$ or zero.

### 22.6.2   VECTOR E PARALLEL TO THE PLANE OF INCIDENCE (TRANSVERSE MAGNETIC CASE)

Assume that the reference directions of the field vectors of the incident, reflected, and refracted waves in this case is as in Fig. 22.9. Again let $E_{1i}$, $E_{1r}$, and $E_2$ and $H_{1i}$, $H_{1r}$, and $H_2$ be the complex rms values of the field vectors of the three waves at $z = 0$. The boundary conditions in this case are

$$(E_{1i} - E_{1r}) \cos \theta_1 = E_2 \cos \theta_2 \qquad H_{1i} + H_{1r} = H_2. \qquad (22.40)$$

Expressing the magnetic field intensities as $E/\eta$ with appropriate subscripts, we again obtain two linear equations in unknowns $E_{1r}$ and $E_2$. The solution of these equations is straightforward. The reflection and transmission coefficients are found to be
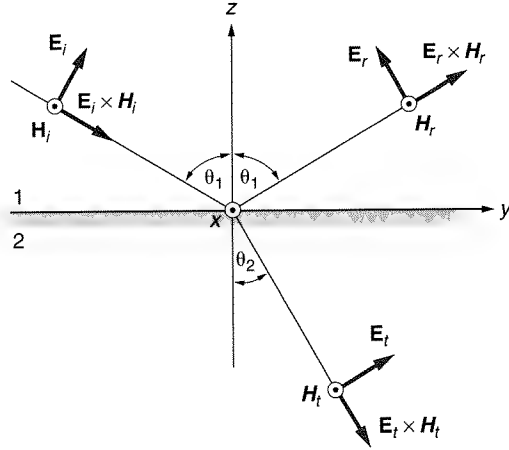
Figure 22.9 Reference directions of the field
vectors of the incident, reflected, and refracted
waves for the parallel polarization of the
incident wave

$$\rho_p = \left(\frac{E_{1r}}{E_{1i}}\right)_p = \frac{\eta_1 \cos\theta_1 - \eta_2 \cos\theta_2}{\eta_1 \cos\theta_1 + \eta_2 \cos\theta_2}, \tag{22.41}$$

$$\tau_p = \left(\frac{E_2}{E_{1i}}\right)_p = \frac{2\eta_2 \cos\theta_1}{\eta_1 \cos\theta_1 + \eta_2 \cos\theta_2}. \tag{22.42}$$

*[Fresnel's coefficients for parallel (TM) polarization]*

Of course, in these two expressions $\cos\theta_2$ must also be calculated as in Eq. (22.39). The coefficients $\rho_p$ and $\tau_p$ in Eqs. (22.41) and (22.42) are the parallel polarization Fresnel coefficients, sometimes also called the *transverse magnetic (TM)* Fresnel coefficients.

   **Example 22.7—Transmission-line models for oblique incidence of plane waves.** We mentioned earlier that transmission-line theory can be used for plane waves incident normally to any interface. It turns out that with a slight modification, we can also use transmission-line theory for oblique incidence. This can be seen if we rewrite the Fresnel coefficients, Eqs. (22.37) and (22.42), as
41

$$\rho_n = \frac{\dfrac{\eta_2}{\cos\theta_2} - \dfrac{\eta_1}{\cos\theta_1}}{\dfrac{\eta_2}{\cos\theta_2} + \dfrac{\eta_1}{\cos\theta_1}} = \frac{\eta_{2n} - \eta_{1n}}{\eta_{2n} + \eta_{1n}}, \tag{22.43}$$

$$\rho_p = -\frac{\eta_2 \cos\theta_2 - \eta_1 \cos\theta_1}{\eta_2 \cos\theta_2 + \eta_1 \cos\theta_1} = -\frac{\eta_{2p} - \eta_{1p}}{\eta_{2p} + \eta_{1p}}, \tag{22.44}$$

where we have now defined the normal and parallel wave impedances as $\eta_{in} = \eta_i / \cos \theta_i$ and $\eta_{ip} = \eta_i \cos \theta_i$. (The minus sign in $\rho_p$ results from the adopted reference directions, and is of no importance.) The transmission coefficients can, obviously, be written in the same way. As an exercise, it is suggested that the reader determine $\rho$ for a normally polarized wave incident at a 45-degree angle from air on a dielectric with $\epsilon_r = 4$ and $\mu = \mu_0$.

**Example 22.8—Fresnel coefficients for perfect dielectrics with equal permeabilities.** In practice the most common case is actually the special case of the two media being perfect dielectrics of equal permeabilities. Then $\eta_1/\eta_2 = \sqrt{\epsilon_2/\epsilon_1}$, and the reflection and transmission coefficients for the normal polarization in Eqs. (22.37) and (22.38) become

$$\rho_n = \frac{\cos \theta_1 - \sqrt{\epsilon_2/\epsilon_1} \cos \theta_2}{\cos \theta_1 + \sqrt{\epsilon_2/\epsilon_1} \cos \theta_2}, \qquad \tau_n = \frac{2 \cos \theta_1}{\cos \theta_1 + \sqrt{\epsilon_2/\epsilon_1} \cos \theta_2} \qquad (\mu_1 = \mu_2). \qquad (22.45)$$

For the parallel polarization, the reflection and transmission coefficients in this case become

$$\rho_p = \frac{\sqrt{\epsilon_2/\epsilon_1} \cos \theta_1 - \cos \theta_2}{\sqrt{\epsilon_2/\epsilon_1} \cos \theta_1 + \cos \theta_2}, \qquad \tau_p = \frac{2 \cos \theta_1}{\sqrt{\epsilon_2/\epsilon_1} \cos \theta_1 + \cos \theta_2} \qquad (\mu_1 = \mu_2). \qquad (22.46)$$

**Example 22.9—The Brewster angle.** From Example 22.8, a few simple conclusions can be drawn:

1. It is not difficult to understand that $\rho_n$ can never be zero. This would require that, simultaneously, $\sin \theta_1 / \sin \theta_2 = \sqrt{\epsilon_2/\epsilon_1}$ (Snell's law) and $\cos \theta_1 / \cos \theta_2 = \sqrt{\epsilon_2/\epsilon_1}$ (the equation resulting from $\rho_n = 0$), which is not possible.

2. If $\theta_1$ is greater than the critical angle for total reflection, we know that $\sin^2 \theta_1 > \epsilon_2/\epsilon_1$, so that $\cos \theta_2$ is purely imaginary. We see from Eq. (22.45) that $\rho_n$ is then in the form $(a - jb)/(a + jb)$. This means that the magnitude of $\rho_n$ is equal to one, that is, that the entire energy of the incident wave is reflected back into medium 1. The same conclusion can be reached for $\rho_p$.

3. The reflection coefficient in the parallel polarization case *can* be zero. For that to happen, it is necessary that $\cos \theta_1 / \cos \theta_2 = \sqrt{\epsilon_1/\epsilon_2}$. This is now not in contradiction with Snell's law, but both equations must simultaneously be satisfied. If we multiply the two equations, we obtain that the reflected wave does not exist if

$$\sin \theta_1 \cos \theta_1 = \sin \theta_2 \cos \theta_2, \qquad \text{or} \qquad \sin 2\theta_1 = \sin 2\theta_2. \qquad (22.47)$$

This equation is satisfied if $2\theta_1 = (\pi - 2\theta_2)$, that is, if $(\theta_1 + \theta_2) = \pi/2$. But for two angles adding to $\pi/2$ the sine of one equals the cosine of the other, so that from Snell's law,

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{\sin \theta_1}{\cos \theta_1} = \tan \theta_1 = n_{12}. \qquad (22.48)$$

*(The Brewster angle, parallel polarization only)*

This particular angle of incidence of a wave with parallel polarization, for which the reflected wave disappears, is known as the *Brewster angle* or the *polarization angle*.

**Example 22.10—Polarization of reflected waves incident at the Brewster angle.** We know that an arbitrarily polarized wave can always be represented as a superposition of two

linearly polarized waves. Therefore, if we have, for example, an elliptically polarized wave incident at the Brewster angle, the reflected wave will not contain the component with parallel polarization, i.e., it will have only a normally polarized electric field. In other words, any wave incident on a plane interface of two dielectric media at the Brewster angle will be reflected as a *linearly polarized wave*.

**Example 22.11—Elimination of the reflected wave in the case of an arbitrarily polarized incident wave.** Suppose that we introduce in the path of the reflected wave from the preceding example a dielectric slab oriented so that the wave is incident on it at the Brewster angle, and that the polarization of the wave (recall that it is defined with respect to the plane of incidence) is parallel. The reflected wave is then going to disappear completely. This is exactly how this phenomenon was discovered experimentally by Brewster, using electromagnetic waves in the visible light frequency region.

*Questions and problems:*    Q22.9 and Q22.10, P22.15 to P22.19

## 22.7   Chapter Summary

1. If a plane wave is incident on a plane boundary surface between two media, boundary conditions can be satisfied by assuming that the wave resulting from the discontinuity (the scattered wave) consists of a reflected plane wave, and (if the other medium is not perfectly conducting) of a transmitted, or refracted, plane wave. This enables relatively simple analysis of electromagnetic scattering of plane waves, similar to transmission-line analysis.

2. If a plane wave is normally incident on a perfectly conducting plane, a standing wave in front of the plane results. If incidence is not normal, there is a standing wave in the direction normal to the plane, and a traveling wave parallel to it.

3. If a plane wave is incident on a plane boundary surface between two dielectric media, a plane wave is reflected from the interface, and a plane wave is transmitted into the other medium.

4. For an arbitrary angle of the incident wave, the plane of incidence is defined as the plane normal to the boundary and containing the direction of propagation of the incident wave.

5. The incident wave is said to have normal polarization if **E** is normal to the plane of incidence, and to have parallel polarization if **E** is parallel to that plane.

6. The ratios of the amplitudes of the reflected and incident waves, and of the transmitted and incident waves, are known as the Fresnel coefficients, with one set for normal polarization and one for parallel polarization.

## QUESTIONS

**Q22.1.** For what orientation and position of a small wire loop, Fig. 22.1, is the emf induced in it maximal?

**Q22.2.** Prove that the time-average value of the Poynting vector at any point in Fig. 22.1 is zero.

**Q22.3.** If waves are represented in phasor form, how can you distinguish a standing wave from a traveling wave?

**Q22.4.** If waves are represented in the time domain, how can you distinguish a standing wave from a traveling wave?

**Q22.5.** Does the emf induced in a small loop of area $S$ placed in Fig. 22.3 at a coordinate $z > 0$ depend on $z$? Does it depend on $z$ if $z < 0$? Explain.

**Q22.6.** Can a small wire loop be placed in Fig. 22.5 so that the emf induced in it is practically zero irrespective of the orientation of the loop?

**Q22.7.** Repeat question Q22.6 for Fig. 22.6.

**Q22.8.** Is total reflection possible if a wave is incident from air onto a dielectric surface? Explain.

**Q22.9.** Why is there no counterpart of the Brewster angle for a wave with vector **E** normal to the plane of incidence?

**Q22.10.** A linearly polarized plane wave is incident from air onto a dielectric half-space, with the vector **E** at an angle $\alpha\,(0 < \alpha < \pi/2)$ with respect to the plane of incidence. Is the polarization of the transmitted and reflected wave linear? If not, what is the polarization of the two waves? Does it depend, for a given $\alpha$, on the properties of the dielectric medium?

## PROBLEMS

**P22.1.** A linearly polarized plane wave of rms electric field strength $E$ and angular frequency $\omega$ is normally incident from a vacuum on a large, perfectly conducting flat sheet. Determine the induced surface charges and currents on the sheet.

**P22.2.** Note that the induced surface currents in problem P22.1 are situated in the magnetic field of the incident wave. Determine the time-average force per unit area (the pressure) on the sheet. (This is known as *radiation pressure*.)

**P22.3.** Repeat problems P22.1 and P22.2 assuming the wave is polarized circularly.

**P22.4.** If the conductivity $\sigma$ of the sheet in problem P22.1 is large, but finite, its permeability is $\mu$, and the frequency of the wave is $\omega$, find the time-average power losses in the sheet per unit area. Specifically, find these losses if $f = 1\,\text{MHz}$, $E = 1\,\text{V/m}$, $\sigma = 56 \cdot 10^6\,\text{S/m}$ (copper), and $\mu = \mu_0$.

**P22.5.** A plane wave, of wavelength $\lambda$, is normally incident from a vacuum on a large, perfectly conducting sheet. A circular loop of radius $a(a \ll \lambda)$ should be at a location at which the induced emf is maximal, as near as possible to the sheet. If the electric field of the incident wave is $E$, calculate this maximal emf.

**P22.6.** Assume that a time-harmonic surface current of density $J_{sx} = J_{s0} \cos \omega t$ exists over an infinitely large plane sheet. Write the integral expression for the electric field strength vector at a distance $z$ from the sheet. Do not evaluate the integral, but reconsider problem P22.1 to see if you know what the result must be.

**P22.7.** A linearly polarized plane wave, of frequency $f = 1\,\text{MHz}$, is normally incident from a vacuum on the planar surface of distilled water ($\mu = \mu_0$, $\epsilon = 81\epsilon_0$, $\sigma \simeq 0$). The rms value of the electric field strength of the incident wave is $E = 100\,\text{mV/m}$. A loop of area $S = 100\,\text{cm}^2$ wound with $N = 5$ turns is situated in water so that the emf induced in it is maximal. Determine the rms value of the emf.

**P22.8.** A plane wave propagating in dielectric 1, of permittivity $\epsilon_1$ and permeability $\mu_1$, impinges normally on a dielectric slab 2, of permittivity $\epsilon_2$, permeability $\mu_2$, and thickness $d$. To the right of the slab there is a semi-infinite medium of permittivity $\epsilon_3$ and permeability $\mu_3$. Determine the reflection coefficient at the interface between media 1 and 2. Plot the reflection coefficient as a function of the slab thickness, $d$, for given permittivities. Consider cases when (1) $\epsilon_1 > \epsilon_2 > \epsilon_3$, (2) $\epsilon_3 > \epsilon_2 > \epsilon_1$, (3) $\epsilon_2 > \epsilon_1 > \epsilon_3$, and (4) $\epsilon_2 > \epsilon_3 > \epsilon_1$.

**P22.9.** Assume that in the preceding problem the thickness of the slab is (1) half a wavelength, and (2) a quarter of a wavelength in the slab. Determine the relationship between the intrinsic impedances of the three media for which in the two cases there will be no reflected wave into medium 1. (The first of these conditions is used in antenna covers, called radomes. The second is used in optics, for so-called anti-reflection, or AR, coatings. The thickness and relative permittivity of a thin transparent layer over lenses can be designed in this way so that the reflection of light from the lens is minimized.)

**P22.10.** Find the reflection and transmission coefficients for the interface between air and fresh water ($\epsilon = 81\epsilon_0, \sigma \simeq 0$), in the case of perpendicular incidence.

**P22.11.** A plane wave is normally incident on the interface between air and a dielectric having a permeability $\mu = \mu_0$, and an unknown permittivity $\epsilon$. The measured standing-wave ratio in air is 1.8. Determine $\epsilon$.

**P22.12.** What is the position of a small loop of area $S$ in Fig. 22.6 in order that the emf induced in it be maximal? If the electric field of the wave is $E$ and its frequency $f$, calculate this maximal emf.

**P22.13.** Repeat problem P22.12 for a small loop placed in the wave in Fig. 22.5.

**P22.14.** Determine the minimal relative permittivity of a dielectric medium for which the critical angle of total reflection from the dielectric into air is less than 45 degrees. Is it possible to make from such a dielectric a right-angled isosceles triangular prism that returns the light wave as in Fig. P22.14? Is there reflection of the light wave when it enters the prism?
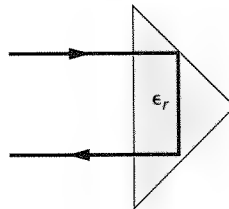


Figure P22.14 Reflection of a light wave by a prism

**P22.15.** A plane wave with parallel polarization is incident at an angle of $\pi/4$ from air on a perfect dielectric with $\epsilon_r = 4$ and $\mu = \mu_0$. Find the Fresnel coefficients. What fraction of the incident power is reflected, and what is transmitted into the dielectric? More generally, plot the Fresnel coefficients and the reflected and transmitted power as a function of $\epsilon_r$, assuming its value is between 1 and 80.

**P22.16.** Repeat problem P22.15 for a normally polarized wave.

**P22.17.** A plane wave with normal polarization is incident at an angle of 60° from air onto deep fresh water with $\epsilon_r = 81$ ($\sigma = 0$). The rms value of the incident electric field is 1 V/m. Find the rms value of the reflected and transmitted electric field.

**P22.18.** Repeat problem P22.17 for parallel polarization.

**P22.19.** Is there an incident angle in problems P22.17 and P22.18 for which the reflected wave is eliminated? If so, calculate this angle for the two polarizations.

# 23

# *Waveguides and Resonators*

## 23.1 Introduction

Waveguides are structures that direct electromagnetic energy along a desired path, transmission lines being just one example. We know that transmission lines consist of two conductors, but some may have more than two, as in three-phase power lines. Maxwell's equations predict that electromagnetic waves can also be guided through hollow metallic tubes, like water is "guided" through pipes. There are a variety of such hollow metallic waveguides, differing in the shape of their cross section; the most common shape is rectangular.

Maxwell's equations also predict that electromagnetic waves can be guided by dielectric slabs or rods, known as *dielectric waveguides*. For example, an optical fiber is a specific type of dielectric waveguide used for guiding electromagnetic waves at optical frequencies.

Transmission lines can support waves with vectors **E** and **H** in planes *transversal* (normal) to the direction of wave propagation. We know that such waves are termed *transverse electromagnetic waves*, or TEM waves. We already know that plane waves are also TEM waves, but for a plane wave the vectors **E** and **H** in transversal planes are *constant*, whereas in transmission lines they are not.

Waveguides in the form of metallic tubes and dielectric plates or rods cannot support TEM waves. Instead, waves along such waveguides may have *either* the **E** vector *or* the **H** vector in the transversal plane alone, but the other vector must have

432

a component in the direction of propagation. These two wave types are called *transverse electric*, or TE, waves, and *transverse magnetic*, or TM, waves.

We know that lossless transmission lines guide TEM waves of *any* frequency, and *with the same velocity*. We will see that TE and TM waves can propagate only above a certain critical frequency, and that their velocity depends on frequency. So structures supporting TE and TM waves behave as high-pass filters.

We have seen that among other purposes, transmission lines are used as circuit elements (to obtain an element with desired reactance, to act as a transformer, etc.). Waveguides are also used for such purposes, but only in the microwave range because they would be very large and impractical at low frequencies. They are used as building blocks of various microwave components: attenuators, phase shifters, transformers, and so on. We will consider only one such component, the so-called resonant cavity, which is an analogue to an *LC* resonant circuit with a lumped inductor and a lumped capacitor. A resonant cavity, however, is a *spatial* resonator, in the form of a box in which electromagnetic energy oscillates, similar to the way acoustic energy oscillates in a hallway.

The theory of waveguides is significantly more complex than any theory we have considered so far, and a complete presentatoin is beyond the scope of this introductory text. Since the waveguides are of great practical importance at higher frequencies, every electrical engineer should know at least the basic concepts of these electromagnetic structures. A compromise is therefore made in what follows, and most of the basic waveguide equations are given without proof. The interested reader can find these proofs in Appendix 8.

## 23.2   Wave Types (Modes)

Consider a hollow, perfectly conducting waveguide pipe, filled with a perfect dielectric of parameters $\epsilon$ and $\mu$, as in Fig. 23.1. Let the complex vectors **E** and **H** in the waveguide be of the form $\mathbf{E}_{tot} = \mathbf{E}(x, y)e^{-\gamma z}$, and $\mathbf{H}_{tot} = \mathbf{H}(x, y)e^{-\gamma z}$. Here $\gamma$ is the propagation coefficient in the $z$ direction. First we allow $\gamma$ to be complex, and later we will discuss what that physically means. After performing vector differentiation
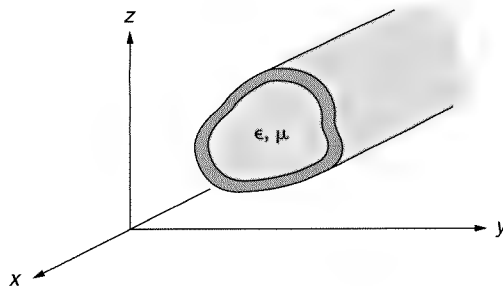


Figure 23.1 Cross section of a general waveguide. It is assumed that the waveguide is lossless, and that the dielectric is homogeneous.

on Maxwell's equations in complex form, as shown in detail in Appendix 8, the following expressions for the electric and magnetic field components are obtained:

$$E_x = -\frac{1}{K^2}\left(\gamma\frac{\partial E_z}{\partial x} + j\omega\mu\frac{\partial H_z}{\partial y}\right), \tag{23.1}$$

$$E_y = -\frac{1}{K^2}\left(\gamma\frac{\partial E_z}{\partial y} - j\omega\mu\frac{\partial H_z}{\partial x}\right), \tag{23.2}$$

$$H_x = -\frac{1}{K^2}\left(-j\omega\epsilon\frac{\partial E_z}{\partial y} + \gamma\frac{\partial H_z}{\partial x}\right), \tag{23.3}$$

$$H_y = -\frac{1}{K^2}\left(j\omega\epsilon\frac{\partial E_z}{\partial x} + \gamma\frac{\partial H_z}{\partial y}\right), \tag{23.4}$$

where

$$K^2 = \gamma^2 + \beta^2, \qquad \beta^2 = \omega^2\epsilon\mu. \tag{23.5}$$

That Eqs. (23.1) to (23.5) are solutions to Maxwell's equations can be checked by substitution. Note that the propagation coefficient $\gamma$ (and therefore also the coefficient $K$) *is not known*. So there are seven scalar unknowns (the six field components and $\gamma$).

We can reach an interesting conclusion by looking carefully at the preceding equations: if we can find $E_z(x, y)$ and $H_z(x, y)$, we know the complete electromagnetic field for a given waveguide shape and size. [Thus the functions $E_z(x, y)$ and $H_z(x, y)$ play a role analogous to a potential function because all the other components are obtained from them by differentiation.]

These equations have three *classes* of solution:

1. Both $E_z = 0$ and $H_z = 0$, that is, only transversal components of the wave exist. [The possibility of such a solution is not evident from Eqs. (23.1) to (23.5), but will be demonstrated in the next section.] This solution corresponds to a TEM wave.
2. $E_z = 0$, but $H_z \neq 0$. This solution corresponds to a transverse electric (TE) wave.
3. $E_z \neq 0$, and $H_z = 0$. This solution corresponds to a transverse magnetic (TM) wave.

We now examine these three classes of solutions, often called *modes*, in turn.

### 23.2.1    TRANSVERSE ELECTROMAGNETIC (TEM) WAVES

The salient properties of TEM wave types, or TEM modes, are the TEM propagation coefficient $\gamma$, the wave impedance $Z_{TEM}$, and the unique quasi-static nature of TEM wave types.

**Propagation Coefficient**

If both $E_z = 0$ and $H_z = 0$, the expressions in parentheses in Eqs. (23.1) to (23.4) are zero. One might be tempted to conclude that all the other components are also zero.

Note, however, that *K is not known*, so it can also be zero. We know that the expression of the form 0/0 need not be undefined (for example, $\sin x/x \to 1$ if $x \to 0$). So the solution could exist only if $K^2 = \gamma^2 + \beta^2 = 0$, or

$$\gamma = \pm j\omega\sqrt{\epsilon\mu}. \tag{23.6}$$

*(Propagation coefficient of TEM waves)*

We recognize in $\gamma$ the propagation coefficient of plane waves, and also the propagation coefficient along lossless transmission lines. We will see that indeed, waves propagating along lossless transmission lines *are* of the TEM type.

## Wave Impedance

In Eqs. (23.1) and (23.4) let $\gamma = \pm j\omega\sqrt{\epsilon\mu}$. After simple manipulations we find that in such a case, the ratio of the transverse electric and magnetic field components is

$$\frac{E_x}{H_y} = \pm Z_{\text{TEM}}, \quad \frac{E_y}{H_x} = \mp Z_{\text{TEM}}, \quad \text{where} \quad Z_{\text{TEM}} = \sqrt{\frac{\mu}{\epsilon}} \tag{23.7}$$

*(Wave impedance of TEM waves)*

for any $E_z$ and $H_z$ (which cancel out). $Z_{\text{TEM}}$ is known as the *wave impedance of TEM waves*.

From this we can draw three conclusions. First, vector **H** is normal to vector **E** (both are, of course, in a transverse plane). Second, the ratio of the electric and magnetic fields for a forward wave (the upper sign) is the intrinsic impedance of the medium, and for the backward wave it is the negative of that. Third, for the forward and for the backward waves **E** and **H** are such that their cross product results in the Poynting vector (power flow) in the respective direction. How do these properties compare to those of a plane wave in free space?

## Quasi-Static Nature of TEM Waves

There is an interesting general conclusion about TEM wave types. It turns out (see Appendix 8, section A8.2) that the electric field is derivable from a potential function which at $z = 0$ satisfies Laplace's equation in $x$ and $y$:

$$\frac{\partial^2 V(x, y)}{\partial x^2} + \frac{\partial^2 V(x, y)}{\partial y^2} = 0. \tag{23.8}$$

Because boundary conditions require that the tangential **E** on conductor surfaces be zero, we reach the following conclusion: for TEM waves, the electric field in planes where $z$ is constant is the same as the electrostatic field corresponding to the potentials of waveguide conductors at that cross section.

**Example 23.1—A TEM wave cannot propagate through a hollow metal tube.** Consider a waveguide in the form of a hollow metal tube. For a TEM wave to exist inside the tube, the field must be the same as in the electrostatic case. But we know that inside a hollow conductor

with no charges there can be no electrostatic field. Consequently, TEM waves cannot propagate through hollow metallic waveguides.

Note that a coaxial cable does have another conductor in the tube, and that an electrostatic field can exist in the cable if the two cable conductors are at different potentials. Therefore, a TEM wave can propagate inside a coaxial cable (which we already know is true).

**Example 23.2—Transmission lines must have at least two conductors.** For the electrostatic field to exist in a cylindrical system, we must have at least two conductors. Indeed, a single charged conductor is a fiction—it implies infinite electrical energy per unit length, since the potential of the conductor with respect to a reference point at infinity is infinite. So TEM waves cannot propagate along a single wire.

However, any electrostatic system of two or more conductors *with a zero total charge per unit length* is feasible, because it has a finite electrical energy per unit length. Consequently, TEM waves can propagate along such waveguides. Note that this also implies a zero total current at any cross section of a transmission line, a proof of which is left as an exercise for the reader. Thus, equations of TEM waves propagating along waveguides are actually equations of wave propagation along lossless transmission lines.

### 23.2.2 TRANSVERSE ELECTRIC (TE) WAVES

Now let us briefly examine the general properties of TE wave types (for details, see Appendix 8, section A8.3).

#### Propagation Coefficient

Using the condition $E_z(x, y) = 0$, the wave equation for the $H_z$ component in this case is given by

$$\frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + \gamma^2 H_z + \omega^2 \epsilon \mu H_z = \frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + K^2 H_z = 0. \tag{23.9}$$

To solve this equation we need to know the geometry of the waveguide. We will see that for specific boundary conditions this equation can be satisfied only for certain distinct values of the parameter $K$. These values of $K$, for which both the wave (Helmholtz) equation and boundary conditions are satisfied, are known as its *eigenvalues*, or *characteristic values*. We will see that, for example, eigenvalues of $K$ for a rectangular waveguide are given by a double infinite set of pairs of numbers dependent on the waveguide dimensions and frequency. An eigenvalue of $K$ determines the propagation coefficient $\gamma$ according to Eq. (23.5).

#### Wave Impedance

From Eqs. (23.1) to (23.4) it follows that if $E_z = 0$, the *transverse electric and magnetic field vectors in a TE wave are normal to each other*, and that

$$Z_{\text{TE}} = \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{j\omega\mu}{\gamma} \tag{23.10}$$

*(Wave impedance of TE waves)*

is a constant, the same at all points of the field in a waveguide. This is known as the *wave impedance of TE modes*. We will see that it is *not* the same as that for TEM waves, because $\gamma$ for TE waves is different from $j\omega\sqrt{\epsilon\mu}$.

### 23.2.3  TRANSVERSE MAGNETIC (TM) WAVES

Finally, let us look at the general properties of TM wave types.

#### Propagation Coefficient

Using the condition $H_z(x, y) = 0$, we can obtain $E_z$ from the Helmholtz equation, which now reads

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + K^2 E_z = 0. \tag{23.11}$$

To solve this equation we need to know the geometry of the waveguide, and a solution exists only for specific values of $K$ (its eigenvalues), as in the case of TE modes.

#### Wave Impedance

As in the case of TE modes, from Eqs. (23.1) to (23.4) it follows that for $H_z = 0$, the transverse electric and magnetic field vectors in a TM wave are normal to each other, and that

$$\frac{E_x}{H_y} = -\frac{E_y}{H_x} = Z_{TM} = \frac{\gamma}{j\omega\epsilon} \tag{23.12}$$

*(Wave impedance of TM waves)*

is the same at all points. This is known as the *wave impedance of TM wave types*.

Note that for a forward wave and the same value of the propagation coefficient $\gamma$,

$$Z_{TE}Z_{TM} = Z_{TEM}^2 = \frac{\mu}{\epsilon}. \tag{23.13}$$

**Example 23.3—Power transmitted along waveguides.** Let us now derive a general expression for the power transmitted along a waveguide. Let only a forward wave exist in a waveguide. The power transmitted along the waveguide can be determined by integrating the complex Poynting vector over the structure cross section at $z = 0$:

$$P = \int_{S_{transv}} \text{Re}\{(\mathbf{E}_{transv} \times \mathbf{H}^*_{transv}) \cdot \mathbf{u}_z\} \, dS_{transv}, \tag{23.14}$$

where the subscript "transv" relates to components normal to the direction of propagation.

The transverse components of vectors $\mathbf{E}$ and $\mathbf{H}$ for all three wavetypes (TEM, TE, and TM) are normal to each other. In addition, their ratio equals the wave impedance of the wave,

and the cross product $\mathbf{E}_{\text{transv}} \times \mathbf{H}^*_{\text{transv}}$ is in the direction of vector $\mathbf{u}_z$. So

$$(\mathbf{E}_{\text{transv}} \times \mathbf{H}^*_{\text{transv}}) \cdot \mathbf{u}_z = E_{\text{transv}} H^*_{\text{transv}} = Z_{\text{wave type}} |H_{\text{transv}}|^2 = \frac{1}{Z_{\text{wave type}}} |E_{\text{transv}}|^2,$$

where $Z_{\text{wave type}}$ is the wave impedance of the wave propagating along the waveguide. Thus for the power transmitted along the waveguide we obtain

$$P_{\text{wave type}} = Z_{\text{wave type}} \int_{S_{\text{transv}}} |H_{\text{transv}}|^2 dS_{\text{transv}}$$

$$= \frac{1}{Z_{\text{wave type}}} \int_{S_{\text{transv}}} |E_{\text{transv}}|^2 dS_{\text{transv}} \qquad \text{(valid for forward wave),} \qquad (23.15)$$

where "wave type" stands for TEM, TE, or TM.

*Questions and problems:* Q23.1 to Q23.7, P23.1 to P23.4

## 23.3 Rectangular Metallic Waveguides

At frequencies above about 1 GHz, losses in transmission-line conductors due to skin effect become pronounced, so lower-loss hollow waveguides are used for guiding waves up to frequencies of several hundred gigahertz. Most often such waveguides are of rectangular cross section, but circular and some other cross sections are also used. We restrict our attention to rectangular waveguides.

A sketch of a rectangular waveguide is shown in Fig. 23.2. We assume the waveguide to be lossless and straight. From Example 23.1 we know that TEM waves cannot propagate along hollow waveguides. So we consider TE modes in more detail, and also TM wave types briefly. Let us start with the TE wave types.
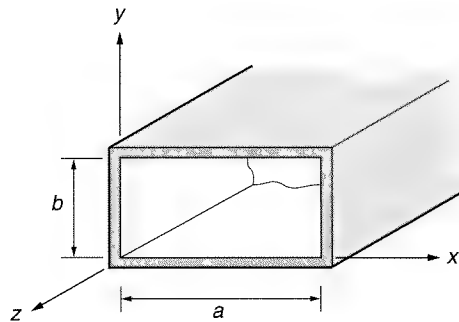


Figure 23.2 Sketch of a waveguide of rectangular cross section

### 23.3.1   TE WAVES IN RECTANGULAR WAVEGUIDES

The complete derivation for TE modes in a rectangular metallic waveguide is given in Appendix 8, section A8.3. At the introductory level, it suffices to quote the most important expressions and discuss their practical meaning.

**Complete Expression for TE$_{mn}$ Wave Types**

After applying the boundary conditions to the perfectly conducting waveguide walls at $x = 0$, $x = a$, $y = 0$, and $y = b$, the $H_z$ component in the cross section $z = 0$ of the waveguide is found to be

$$H_z(x, y) = H_0 \cos\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \qquad \text{(at } z = 0\text{),} \tag{23.16}$$

where $H_0$ is a constant depending on the level of excitation of the wave in the waveguide. The other components at $z = 0$ are (note that $E_z = 0$)

$$E_x(x, y) = \frac{j\omega\mu}{K^2}\frac{n\pi}{b}H_0 \cos\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \qquad \text{(at } z = 0\text{),} \tag{23.17}$$

$$E_y(x, y) = -\frac{j\omega\mu}{K^2}\frac{m\pi}{a}H_0 \sin\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \qquad \text{(at } z = 0\text{),} \tag{23.18}$$

$$H_x(x, y) = \frac{\gamma}{K^2}\frac{m\pi}{a}H_0 \sin\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \qquad \text{(at } z = 0\text{),} \tag{23.19}$$

$$H_y(x, y) = \frac{\gamma}{K^2}\frac{n\pi}{b}H_0 \cos\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \qquad \text{(at } z = 0\text{).} \tag{23.20}$$

Since the cosine and sine functions are periodic, we see that there is a double infinite number of TE wave types, corresponding to any possible pair of $m$ and $n$. Note that $m$ represents the number of half-waves along the $x$ axis, and $n$ the number of half-waves along the $y$ axis. The wave determined by $m$ and $n$ is known as a $TE_{mn}$ *mode*. From Eqs. (23.16) to (23.20) we see that for $m = n = 0$ all components are zero. Thus, a TE$_{00}$ mode does not exist. Waves for any other combinations of numbers $m$ and $n$ may exist, for example, TE$_{10}$, TE$_{01}$, TE$_{11}$, TE$_{21}$. The values of the wave components for any $z$ are obtained by simply multiplying the preceding expressions by $e^{-\gamma z} = e^{-j\beta z}$, where $\gamma$ and $\beta$ are as given in the next section.

**Propagation Coefficient of TE Waves**

Noting that $\omega = 2\pi f$, the expression for the propagation coefficient of a TE wave along a rectangular waveguide is

$$\gamma = j\beta \qquad \beta = \omega\sqrt{\epsilon\mu}\sqrt{1 - \frac{f_c^2}{f^2}}, \tag{23.21}$$

*(Propagation and phase coefficients of rectangular waveguides)*

where $f_c$ is known as the *cutoff frequency*, and it depends on the mode numbers $m$ and $n$, and on the dimensions of the waveguide, $a$ and $b$.

## Cutoff Frequency of TE Wave Types

Noting that $1/\sqrt{\epsilon\mu} = c$,

$$f_c = \frac{c}{2}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}, \qquad c = \frac{1}{\sqrt{\epsilon\mu}}. \tag{23.22}$$

*(Cutoff frequency of rectangular waveguides)*

Why $f_c$ is known as the cutoff frequency is explained next.

## Phase and Group Velocity of TE$_{mn}$ Waves

Let us now investigate more closely the properties of TE$_{mn}$ modes for different pairs of values of $m$ and $n$. First, note that the *phase velocity* of the TE$_{mn}$ mode is given by

$$v_{\text{ph}} = \frac{\omega}{\beta} = \frac{c}{\sqrt{1 - f_c^2/f^2}}. \tag{23.23}$$

*(Phase velocity of waves propagating along rectangular waveguides)*

You may recall Example 21.5, in which we showed that for this dependence of the phase velocity on frequency, the group velocity is given by

$$v_g = c\sqrt{1 - f_c^2/f^2}. \tag{23.24}$$

*(Group velocity of waves propagating along rectangular waveguides)*

Since the phase (and group) velocity depend on frequency, rectangular waveguides are *dispersive structures*. The dependence of the phase and group velocity on normalized frequency, $f/f_c$, is sketched in Fig. 23.3.
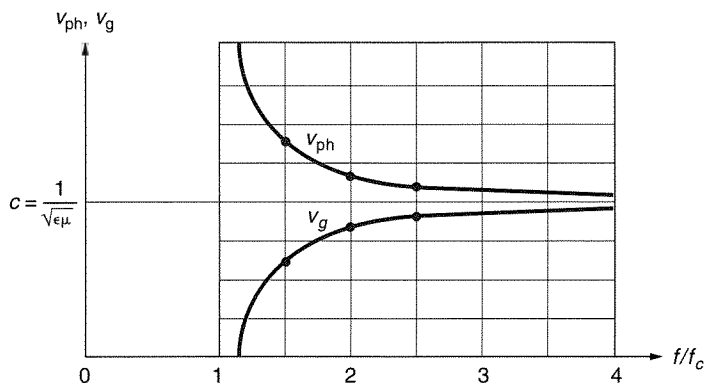


Figure 23.3 Dependence of phase velocity and group velocity on normalized frequency, $f/f_c$
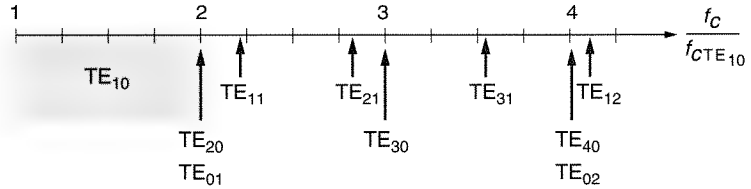
Figure 23.4 Cutoff frequencies of the first few higher-order TE modes normalized to that of the dominant $TE_{10}$ mode, for $a/b = 2$

Assume that for a given waveguide $f > f_c$. Then the phase coefficient is real, as are the phase velocity and the group velocity. This means that the wave of this frequency propagates along the waveguide.

If, however, $f < f_c$, the expression under the square root in Eqs. (23.21), (23.23), and (23.24) is *negative*. The phase coefficient becomes imaginary, $-j|\beta|$ (negative value of the root is taken to avoid exponentially increasing wave amplitudes with increasing $z$), so that the propagation factor $e^{-j\beta z}$ becomes $e^{-|\beta|z}$. *This means that waves of frequencies lower than $f_c$ cannot propagate along rectangular waveguides.* This is why $f_c$ is termed the "cutoff frequency." Because the attenuation of the wave is exponential, the wave of a frequency $f < f_c$ is attenuated very rapidly with $z$. Thus, as mentioned in the introduction to this chapter, rectangular waveguides behave as high-pass filters.

Modes that propagate through a given waveguide are the *propagating modes*, and those that do not propagate are the *evanescent modes*. To transmit energy, we use propagating modes. Evanescent modes are used, for example, when making an attenuator out of a section of a waveguide.

Rectangular waveguides are always made such that $a > b$. Let $a = 2b$, which is fairly standard. The first few cutoff frequencies of the $TE_{mn}$ modes, Eq. (23.22), relative to the cutoff frequency of the $TE_{10}$ mode are shown in Fig. 23.4. Note that between the cutoff frequency of the $TE_{10}$ mode and the next one, that of the $TE_{01}$ mode, only the $TE_{10}$ mode can propagate. This is a remarkable property of the $TE_{10}$ mode. A discontinuity in the waveguide, like a bend or a slot in the guide wall, will always produce a multitude of modes. If we use a frequency of a wave to be within these limits (shown shaded in Fig. 23.4), out of all of these modes only the $TE_{10}$ mode will propagate further—all other modes will be evanescent.

### 23.3.2    TM WAVES IN RECTANGULAR WAVEGUIDES

As in the case of TE modes, there are an infinite number of $TM_{mn}$ modes, corresponding to all possible pairs of numbers $m$ and $n$. The expressions for the cutoff frequency, propagation coefficient, phase velocity, etc., are very similar to the TE case. There is an important difference, however: the lowest TM mode is $TM_{11}$, that is, there is no TM mode for which either $m = 0$ or $n = 0$. As an illustration, Fig. 23.5b shows the comparison of the lowest order TE mode fields and the $TM_{11}$ mode fields.

## 23.4  TE$_{10}$ Mode in Rectangular Waveguides

We have seen that the cutoff frequencies of all TE modes higher than the 10 mode, as well as the cutoff frequencies of all TM modes, are higher than that of the TE$_{10}$ mode. For this reason the TE$_{10}$ mode is known as the *dominant mode* in rectangular waveguides. It is by far the most commonly used wave type in hollow metallic waveguides, so we consider it in more detail.

The field components of the TE$_{10}$ mode are given by Eqs. (23.16) to (23.20) for $m = 1$ and $n = 0$:

$$H_z(x, y) = H_0 \cos\left(\frac{\pi}{a}x\right) \quad \text{(TE$_{10}$ mode)}, \tag{23.25}$$

$$E_x(x, y) = E_z(x, y) = H_y(x, y) = 0 \quad \text{(TE$_{10}$ mode)}, \tag{23.26}$$

$$E_y(x, y) = -\mathrm{j}\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right) \quad \text{(TE$_{10}$ mode)}, \tag{23.27}$$

$$H_x(x, y) = \mathrm{j}\beta\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right) \quad \text{(TE$_{10}$ mode)}. \tag{23.28}$$

*(Field components of* TE$_{10}$ *mode in a rectangular waveguide)*

The cutoff frequency of the TE$_{10}$ mode is

$$f_{c\,\mathrm{TE10}} = \frac{c}{2a}, \tag{23.29}$$

*(Cutoff frequency of* TE$_{10}$ *mode in rectangular waveguide)*

and the phase velocity, wave impedance, and so on for the TE$_{10}$ mode are obtained from the general expressions with this cutoff frequency.

The wave impedance of the TE$_{10}$ mode is obtained from Eqs. (23.10), (23.21), and (23.29):

$$Z_{\mathrm{TE10}} = \frac{\sqrt{\mu/\epsilon}}{\sqrt{1 - f_c^2/f^2}} > \sqrt{\mu/\epsilon}, \tag{23.30}$$

where $f_c = c/2a$. This expression tells us that the field inside the waveguide is different from that in free space (a plane wave). Consequently, if we cut a waveguide, only part of the energy will be radiated from its open end, and the rest will be reflected back.

**Example 23.4—Wavelength along waveguide.** The wavelength inside the waveguide, along the z axis, is determined simply as $\lambda_z = 2\pi/\beta$, where $\beta$ is the phase coefficient of the mode of interest. So the wavelength along the waveguide is

$$\lambda_z = \frac{2\pi}{\beta} = \frac{c}{f\sqrt{1 - f_c^2/f^2}} = \frac{\lambda_0}{\sqrt{1 - f_c^2/f^2}}, \tag{23.31}$$

*(Wavelength along a rectangular waveguide)*

where $\lambda_0$ is the wavelength of a plane wave of the same frequency and in the same medium.

(a)



(b)
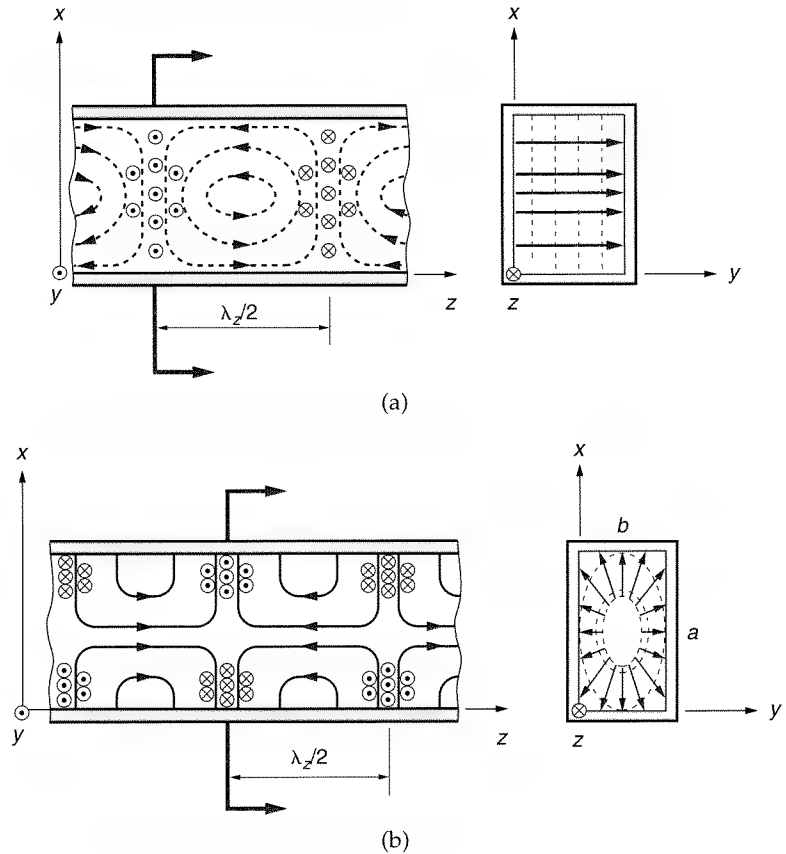
**Figure 23.5** (a) Sketch of the $E$-field and $H$-field lines of the $TE_{10}$ mode in a rectangular waveguide, frozen in time. (b) Sketch of the $E$ and $H$ lines of the $TM_{11}$ mode (the mode with the lowest cutoff frequency of all TM modes). Solid lines show the $E$ lines, and dashed lines and $\odot$ and $\otimes$ symbols show the $H$ lines. The entire picture moves at the phase velocity in the $+z$ direction.

**Example 23.5—Sketch of the field of the $TE_{10}$ mode.** When we multiply Eqs. (23.25) to (23.28) by $e^{j\beta z}$, we obtain the phasor wave components at any point $(x, y, z)$. To obtain a picture of the fields in the waveguide at an instant, we need to obtain the time-domain expressions, fix an instant in time (e.g., $t = 0$), and then plot the field lines. Although time-domain expressions are easy to obtain (this is left as an exercise for the reader), plotting the field is not simple. A sketch of the fields of the $TE_{10}$ mode is shown in Fig. 23.5a. In time, the entire picture moves in the $+z$ direction with the phase velocity of the $TE_{10}$ mode. For comparison, the $E$ and $H$ lines of the $TM_{11}$ mode (the mode with the lowest cutoff frequency of all TM modes) are sketched in Fig. 23.5b.

**Example 23.6—Surface current distribution on waveguide walls for the $TE_{10}$ mode.** The surface currents are obtained from the time-domain expressions of the magnetic field on waveguide walls and the boundary condition in Eq. (19.12) for a perfect conductor. We need to
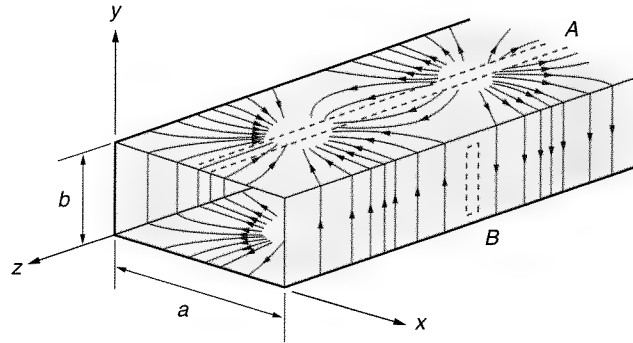
Figure 23.6 Sketch of surface current distribution for the
$TE_{10}$ mode in a rectangular waveguide, frozen in time. The
entire picture moves at the phase velocity in the $+z$
direction.

fix an instant of time (e.g., $t = 0$), and then plot the lines of the surface-current density vector. This, again, is not an easy task. A sketch of the lines of the current density vector is shown in Fig. 23.6. In time, the surface current density distribution moves with the phase velocity of the $TE_{10}$ mode in the $+z$ direction.

Note that if we cut a slot in the waveguide wall, in such a way that the slot is always tangential to the lines of the surface current, only a small disturbance of the wave propagation in the waveguide will result. Therefore we can cut narrow slots of types $A$ and $B$ indicated in the figure without changing the fields in the waveguide. The slot of type $A$ is made to obtain a slotted waveguide used for measurements similar to those done by a slotted coaxial line (Example 18.10).

**Example 23.7—Excitation of $TE_{10}$ mode.** How can we produce a $TE_{10}$ mode? First, we need to close one end of the waveguide, to prevent propagation in both directions. We do this with a metal plate perpendicular to the waveguide, as in Fig. 23.7. In waveguide terminology, this is known as a *shorted waveguide*.

To excite the $TE_{10}$ mode, we can excite either the $E$ field or the $H$ field. The $E$ field can be excited by a small coaxial probe. A short extension of the inner conductor of a coaxial line is inserted into the waveguide, and the outer conductor of the line is connected to the waveguide wall, as in Fig. 23.7. Roughly, the position of the probe should be in the middle of the wave-
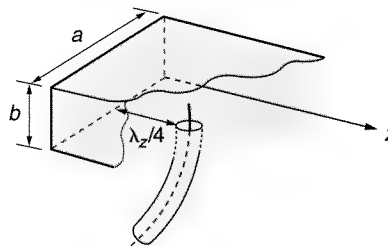


Figure 23.7 Sketch of a probe for exciting a
$TE_{10}$ mode in a rectangular waveguide

guide (where the $E$ field is the strongest, Fig. 23.5), and about a quarter of a guided wavelength $\lambda_z$ from the short circuit. If the latter condition is fulfilled, the wave from the probe needs one quarter of a period to reach the short circuit, is reflected there and changes phase (which is equivalent to losing another two quarters of a period), and then loses another quarter of a period to go back to the probe. So the wave reflected from the short circuit will be in phase with the wave radiated in the $+z$ direction. This simple reasoning, however, is only a rough estimate, and the actual position of the probe needs to be determined either experimentally or using accurate numerical methods.

Another possibility is to excite the $TE_{10}$ mode by a small loop intended to excite the $H$ field at the place where it is the strongest, e.g., in the middle of the waveguide short circuit. It is left as an exercise for the reader to sketch this type of waveguide excitation.

**Example 23.8—$TE_{10}$ mode in X-band waveguide.** The microwave frequency range is divided into so-called bands, and a waveguide of certain dimensions can support waves at frequencies covering the entire band. A commonly used range is the X band (about 8.2 to 12.4 GHz), for which a standard waveguide has $a = 23$ mm and $b = 10$ mm. Using the formulas for $TE_{10}$ modes with these waveguide dimensions, we find that the cutoff frequency is $f_c = 6.52$ GHz. The guided wavelength and impedance are different at different frequencies inside the band. At the center of the band, $f = 10$ GHz, the guided wavelength $\lambda_g = 3.96$ cm, and the impedance $Z_{TE10} = 497\,\Omega$. So, the characteristic impedance is much larger than that of a coaxial cable, which is usually $50\,\Omega$. That means that the probe described in Example 23.7 needs to match the coaxial impedance to that of the waveguide dominant mode.

**Example 23.9—Power transmitted by the $TE_{10}$ mode.** The power transmitted by a forward wave through any waveguide is given in Eq. (23.15). We know the transverse components and the wave impedance of the $TE_{10}$ mode, so we need only to substitute these expressions into Eq. (23.15) and to integrate over the cross-sectional area of the waveguide,

$$P_{TE10} = \frac{\beta}{\omega\mu} \int_{y=0}^{b} \int_{x=0}^{a} \omega^2\mu^2 \frac{a^2}{\pi^2} |H_0|^2 \sin^2\left(\frac{\pi}{a}x\right) dx\,dy = \frac{\omega\mu\beta a^2 |H_0|^2}{\pi^2} b \frac{a}{2} = \frac{\omega\mu\beta a^3 b |H_0|^2}{2\pi^2}.$$

$H_0$ is the rms value of the magnetic field at the magnetic field maximum. If $\beta$ is replaced by its expression in Eq. (23.21), this becomes

$$P_{TE10} = \frac{ab}{2}\sqrt{\frac{\mu}{\epsilon}} \frac{f^2}{f_c^2}\sqrt{1 - \frac{f_c^2}{f^2}}\,|H_0|^2. \tag{23.32}$$

It is very instructive to evaluate this power for a specific case. Let the frequency be $f = 10$ GHz, and let $a = 2$ cm and $b = 1$ cm. The cutoff frequency for this waveguide for the $TE_{10}$ mode is $f_c = c/2a = 3\cdot10^8/0.04 = 7.5$ GHz. Let us calculate the maximal possible power that can be transmitted through this waveguide if it is filled with air of dielectric strength $E_{max} = 30$ kV/cm. The maximal electric field is at the coordinate $x = a/2$. At that point, the electric field *amplitude* has to be less than $E_{max}$. This enables us to calculate the maximal $H_0$ from Eq. (23.27):

$$H_{0max} = \frac{\pi E_{max}/\sqrt{2}}{\omega\mu a}.$$

Substituting this value of $H_0$ into Eq. (23.32), we obtain that the maximal power that can be transmitted is about 800 kW. This is much more than the power that could be transmitted through a coaxial cable or printed line (why?), so waveguides are the guiding medium of choice for high-power applications, such as some radars.

*Questions and problems:*     Q23.8 to Q23.22, P23.5 to P23.10

## 23.5   The Microstrip Line (Hybrid Modes)

We have discussed in detail only one of the TE and briefly one of the TM modes in a rectangular waveguide. There are an infinite number of other TE and TM modes in such guiding structures. However, other structures can support wave types that are a combination of TEM, TE, and TM modes. These wave types are called *hybrid modes*.

As an illustration, we consider a commonly used waveguide, called a *microstrip line*, sketched in Fig. 23.8. A microstrip line is made on a dielectric slab, called the *substrate*. One side of the substrate is coated with metal and acts as the ground electrode, similar to the outer conductor of a coax. A metal strip on the other side of the substrate enables a wave to propagate mostly in the dielectric. The role of the strip is similar to that of the center conductor of a coax. Unlike the coax, however, this structure has an inhomogeneous dielectric. The fields are not contained completely in the dielectric but are partly in air, as sketched in Fig. 23.8. This electric field is usually referred to as a fringing field.

Because there are two conductors in this guide, according to Example 23.2 we may suspect that a microstrip can support a TEM wave. Let us check if this is possible. The boundary condition for the (fringing) tangential electric field tells us that $E_{x,\text{diel}} = E_{x,\text{air}}$. We replace $E_x$ with spatial derivatives of $\mathbf{H}$ from Maxwell's second equation in differential form, $\mathbf{E} = -\mu \partial (\nabla \times \mathbf{H})/\partial t$. Taking into account the boundary condition for the magnetic flux density vector (normal components equal on the two sides of
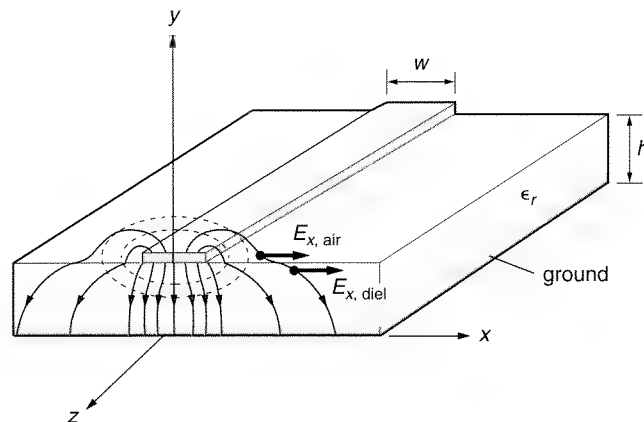


**Figure 23.8** Sketch of a microstrip line. The electric field lines are sketched in solid line, and the magnetic field lines in dashed line.

the boundary), the boundary condition for the electric field can be written as (see problem P23.11)

$$\epsilon_r \frac{\partial H_{z,\text{air}}}{\partial y} - \frac{\partial H_{z,\text{diel}}}{\partial y} = (\epsilon_r - 1)\frac{\partial H_y}{\partial z}. \tag{23.33}$$

Let us see what this boundary condition tells us. The right-hand side is not zero, because $\epsilon_r > 1$, and $H_y$ is not zero. That means that the left-hand side is not zero, which means that $H_z$ is not zero. It can be shown in a similar manner that $E_z$ cannot be zero. So if Maxwell's equations are satisfied for this structure, the wave type that propagates has to have nonzero $H_z$ and $E_z$ components, which means it contains TE and TM modes, and is a hybrid mode.

The components of the electric and magnetic field vectors along the direction of propagation are small compared to the other components, and this structure supports a wave type referred to as a "quasi-TEM" mode, similar to a TEM mode. This means that we define a characteristic impedance and a propagation constant, and then use TEM mode, or transmission-line equations. These line parameters are expressed in terms of an *effective dielectric constant*, which depends on the relative permittivity and thickness of the substrate (see problems P23.12 and P23.13).

Microstrip lines are used extensively at microwave frequencies because of their small size and ease of manufacturing (using printed-circuit board technology). Their loss is higher than that in waveguides, so they are not used for high power levels.

*Questions and problems:*   Q23.23 and Q23.24, P23.11 to P23.13

## 23.6  Electromagnetic Resonators

Classical resonant circuits with lumped elements cannot be used above about 100 MHz. On one hand, losses due to skin effect and dielectric losses become very pronounced. On the other hand, the circuit needs to be sufficiently small not to radiate energy. Therefore at high frequencies, instead of resonant circuits, closed (usually air-filled) metallic structures are used, *inside* which the electromagnetic field is excited to oscillate. Between about 500 MHz and 3 GHz, resonators are usually in the form of shorted segments of shielded transmission lines (e.g., coaxial line, shielded two-wire line). From about 3 GHz to a few tens of GHz, metallic boxes of various shapes are often used instead (most often in the form of a parallelepiped or circular cylinder). Such boxes are known as *cavity resonators*. We have seen in Example 22.1 that at still higher frequencies we use so-called Fabry-Perot resonators, consisting of two parallel, highly polished metal plates.

The basic parameters of an electromagnetic resonator are its *resonant frequency*, $f_r$, the type of wave inside it, and its *quality factor, Q*.

The resonant frequency and the type of wave depend on the resonator shape, size, and excitation, while the quality factor can be defined in general terms. It is defined as the ratio of the electromagnetic energy contained in the resonator, $W_{\text{em}}$, and the total energy lost in one cycle $W_{\text{lost/cycle}}$ in the resonator containing this energy, multiplied by $2\pi$. Since the cycle duration at resonance is $T_r = 1/f_r$, where $f_r$ is the

resonant frequency of the resonator, $W_{\text{lost/cycle}} = P_{\text{losses}} \cdot T_r = P_{\text{losses}}/f_r$. So the $Q$ factor can be written in the following two equivalent forms:

$$Q = 2\pi \frac{W_{\text{em}}}{W_{\text{lost/cycle}}} = \omega_r \frac{W_{\text{em}}}{P_{\text{losses}}} \quad \text{(dimensionless)} \tag{23.34}$$

*(General definition of Q factor of electromagnetic resonators)*

**Example 23.10—Q factor of an *LC* circuit.** The general definition of the $Q$ factor is valid for resonant circuits also. Consider a parallel connection of a capacitor of capacitance $C$ and a coil of inductance $L$. Let the (small) series resistance of the circuit be $R$. Provided that losses are small, we know that the angular resonant frequency of the circuit is $\omega_r = 1/\sqrt{LC}$. We also know that energy oscillates between that in the capacitor and that in the coil. When the energy is completely in the coil, the current in the circuit is maximal, for example, $I_m$. The magnetic energy contained in the coil at that instant is the energy of the circuit, and is simply

$$W_{\text{em}} = \frac{1}{2} L I_m^2.$$

*Time-average* Joule's losses in the circuit corresponding to this current *amplitude* are

$$P_{\text{losses}} = \frac{1}{2} R I_m^2.$$

Thus the circuit $Q$ factor is $Q = \omega_r L/R$, as defined in circuit theory. It is almost impossible to obtain a resonant circuit with a $Q$ factor greater than about 100. (Why do you think this is so? See question Q23.25.)

### 23.6.1 TRANSMISSION-LINE SEGMENTS AS ELECTROMAGNETIC RESONATORS

Consider a two-conductor lossless transmission line segment of length $\zeta$ shorted at its end. We assume the $\zeta$ axis to be directed from the shorted end toward the generator, as in Chapter 18. The input impedance of the segment was derived in Example 18.6:

$$Z(\zeta) = jZ_0 \tan(\beta\zeta) = jZ_0 \tan\left(\frac{2\pi}{\lambda}\zeta\right), \qquad \lambda = \frac{c}{f}, \qquad c = \frac{1}{\sqrt{L'C'}}.$$

We see that $Z(\zeta) = 0$ for $\zeta = n\lambda/2$, $n = 1, 2, \ldots$. This means that a shorted transmission line connected to a generator can be short-circuited at any such point (cross section) without affecting the voltage and current along the line. We can even cut off such a section (shorted at both ends) of the *excited* line, and the current and voltage along it will not be affected. Thus, we obtained an electromagnetic resonator of length $x = n\lambda/2$.

Assume that the rms value of the voltage of the forward wave is $V_+$. The rms value of the forward current wave is $I_+ = V_+/Z_0$. We know from Example 18.3 that the voltage reflection coefficient in this case is $-1$. Therefore, the voltage and current distribution along the line segment, given in Eqs. (18.22), in this case become

$$V(\zeta) = j2V_+ \sin\beta\zeta, \qquad I(\zeta) = 2I_+ \cos\beta\zeta, \tag{23.35}$$

which represent standing voltage and current waves. (Note that in these equations we needed to replace $z$ by $-\zeta$.) We see that the phase difference between the two is $\pi/2$, which means that the voltage is zero everywhere when the current is maximum, and vice versa. Thus we have the same situation as in resonant circuits: when, for example, the current in the resonator is maximal, the entire energy is in the magnetic field. Such resonators can be made with any transmission line, e.g., twin lead, coaxial cable, or microstrip line. The next example discusses a coaxial cable resonator.

**Example 23.11—Q factor of a coaxial resonator.** The energy in a segment $\lambda/2$ of a coaxial transmission line is

$$W_{em} = \int_0^{\lambda/2} \frac{1}{2} L' |I(\zeta)\sqrt{2}|^2 \, d\zeta = \frac{1}{2} L' \frac{8I_+^2 \lambda}{4} = L' I_+^2 \lambda. \tag{23.36}$$

If the conductors have a small resistance $R'$ per unit length, the time-average power losses in the resonator are

$$P_{losses} = \int_0^{\lambda/2} R' |I(\zeta)|^2 \, d\zeta = R' \int_0^{\lambda/2} 4I_+^2 \cos^2(\beta\zeta) \, d\zeta = R' I_+^2 \lambda. \tag{23.37}$$

According to the definition of the $Q$ factor, we obtain that for such a resonator

$$Q = \frac{\omega_r L'}{R'}, \tag{23.38}$$

which is of the same form as for a resonant circuit.

This simple theory is valid for *all* transmission lines. At higher frequencies, however, in open structures like two-wire lines, there may be significant losses due to radiation. Therefore resonators of this type are most often made of a coaxial-line segment, as in Fig. 23.9. The resonator may be excited (and the energy from it extracted) by either a small probe $a$ or a small loop $b$. The probe and the loop are positioned at the voltage maximum (i.e., the electric-field maximum), namely the current maximum (i.e., the magnetic-field maximum).
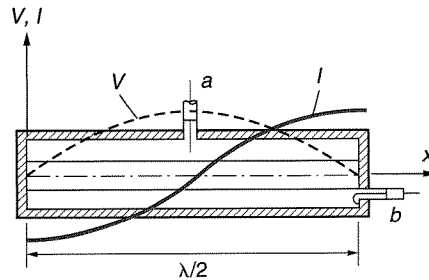


Figure 23.9 Sketch of a resonant ($\lambda/2$) section of a coaxial line shorted at both ends. The resonator may be excited by a small probe $a$, or a small loop $b$. Note the positions of the two excitation elements.

As an example, consider a coaxial resonator with the inner conductor of radius $a =$ 0.5 cm, the inner radius of the outer conductor $b = 1.5$ cm, made of copper ($\sigma = 56 \cdot 10^6$ S/m, $\mu = \mu_0$), at a frequency $f = 1$ GHz. Using the data from Table 18.1 and Eq. (23.38), we obtain that $Q = 4614$. This is a very large value compared with those for resonant circuits (as mentioned, at the most about 100).

### 23.6.2 RESONANT CAVITIES

Resonant cavities may be of diverse shapes, but we will analyze the simplest, in the form of a parallelepiped (rectangular box). It can be obtained by introducing appropriate short circuits (transverse metallic walls) into a rectangular waveguide.

Consider a shorted rectangular waveguide, as in Fig. 23.10. Let a distant generator at left (not shown) excite in the waveguide the dominant, $TE_{10}$ mode. The wave is reflected at the short circuit, giving rise to a backward wave. The backward wave is of the same form as the forward wave, except that the phase coefficient is now $-\beta$. At the short circuit the $E_y$ component of the reflected wave must have the opposite phase with respect to the incident wave. From Eqs. (23.25) to (23.28) we thus obtain the following expressions for the resulting field in the shorted waveguide:

$$E_{y\,\text{res}}(x, y, z) = j\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\left(e^{j\beta z} - e^{-j\beta z}\right)$$

$$= -2\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\sin\beta z, \tag{23.39}$$

$$H_{x\,\text{res}}(x, y, z) = 2j\beta\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\cos\beta z, \tag{23.40}$$

$$H_{z\,\text{res}}(x, y, z) = -2jH_0 \cos\left(\frac{\pi}{a}x\right)\sin\beta z. \tag{23.41}$$

We see that the factor $e^{\pm j\beta z}$ is not present, so we have a standing wave in the waveguide. The other components are zero.
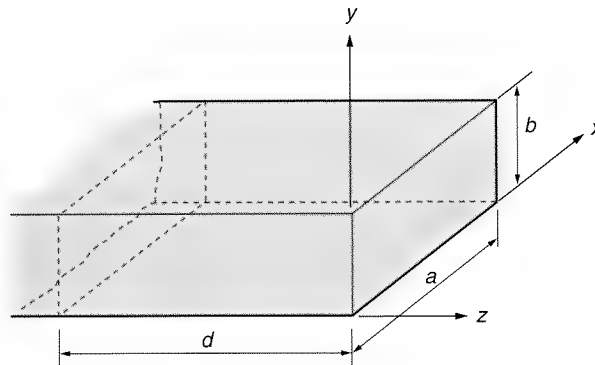


Figure 23.10 Shorted rectangular waveguide. Indicated in dashed lines is another "short circuit" of the waveguide, resulting in a resonant cavity in the form of a rectangular box.

According to Eqs. (23.39) and (23.41), the total $y$ component of the electric field, and the total $z$ component of the magnetic field, are zero not only at $z = 0$ (at the shorted end), but also in planes $z = -\pi p/\beta = -p\lambda_z/2$, $p = 1, 2, \ldots$. Consequently, if we place a thin metal foil in any of these planes and thus short the waveguide at one more place, the field will not change, since the boundary conditions are automatically satisfied in these planes. So we obtain a standing wave in a rectangular box of sides $a$, $b$, and $p\lambda_z/2$. This type of standing wave is known as the $TE_{10p}$ mode in the cavity.

### The $TE_{101}$ Mode

Let us consider the simplest mode, the $TE_{101}$ mode. Let $\lambda_z/2 = d$ in Fig. 23.10. We then have $\beta = 2\pi/\lambda_z = \pi/d$, so that Eqs. (23.39) to (23.41) become

$$E_{y\ res}(x, y, z) = -2\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\sin\left(\frac{\pi}{d}z\right), \tag{23.42}$$

$$H_{x\ res}(x, y, z) = 2j\frac{a}{d}H_0 \sin\left(\frac{\pi}{a}x\right)\cos\left(\frac{\pi}{d}z\right), \tag{23.43}$$

$$H_{z\ res}(x, y, z) = -2jH_0 \cos\left(\frac{\pi}{a}x\right)\sin\left(\frac{\pi}{d}z\right). \tag{23.44}$$

From these equations we can deduce how electromagnetic oscillations in the cavity are maintained. There are induced electric charges on the upper and lower cavity walls because the normal component $E_{y\ res}$ of the electric field vector exists there. Surface currents in the $y$ direction appear only on the side walls, where $H_{x\ res}$ and $H_{z\ res}$ are nonzero. So the oscillations of the electromagnetic field inside the cavity are accompanied by charges and currents on its walls.

According to Eqs. (23.42) to (23.44), the electric and magnetic fields are shifted in phase by $\pi/2$ (the factor j). Therefore, at some points in time there are no currents flowing in the walls, and at others there are no charges. These instants are separated in time by $T/4$, where $T = 1/f$ is the period of the oscillations. Figure 23.11 shows the distribution of charges and currents during one period, starting at the instant when the upper face carries the maximum positive charge.

To determine the quality factor of the cavity, we need to calculate the energy contained in the cavity and the time-average power of losses in the cavity walls corresponding to this energy. At this introductory level, we will just give a numerical example: for a cubic cavity ($a = b = d$) filled with air, designed to operate at 3 GHz in the $TE_{101}$ mode, we find that the side length of the cube is $a = c/(f\sqrt{2}) = 7.07$ cm. (Of course, in that case the $TE_{101}$ mode will be the same as the $TE_{110}$ or $TE_{011}$ modes.) Let the cavity be made of copper ($\sigma = 56 \cdot 10^6$ S/m, $\mu = \mu_0$). Using the losses in the surface resistance of the cavity walls, one can calculate that $Q \simeq 19{,}200$. To obtain this value, the surface resistance is calculated assuming a perfectly polished wall, i.e., with unevenness much less than the skin depth. So as frequency increases, it becomes more and more difficult to avoid increases in loss.

*Questions and problems:*    Q23.25 to Q23.32, and P23.14 to P23.16
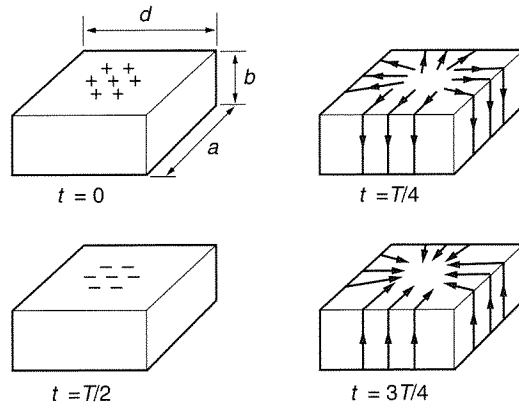
Figure 23.11 Sketch of charge distribution (plus and minus signs) and surface currents over the cavity walls of a rectangular cavity at four different moments in time

## 23.7 Chapter Summary

1. Electromagnetic waves can be guided along a desired route not only by transmission lines but also by hollow pipes, dielectric-coated surfaces, or dielectric rods.

2. It is typical of all these hollow waveguides that (1) they cannot support the TEM wave type, but support the TE and TM wave types, and (2) they cannot transmit energy below a certain frequency, known as the cutoff frequency.

3. Waves of frequencies above the cutoff frequency propagate without attenuation (except due to losses in the materials).

4. Waves of frequencies lower than the cutoff frequency, known as evanescent modes, are exponentially attenuated, and do not propagate at all.

5. Among hollow metallic waveguides, the most important are those of rectangular cross section. There are an infinite number of both $TE_{mn}$ and $TM_{mn}$ modes that can propagate along such waveguides.

6. The higher-order modes (with larger $m$ and $n$ values) for the same waveguide must be of increasingly higher frequencies. So there is a frequency range in which only one mode, the $TE_{10}$ mode, can propagate. This mode is therefore termed the *dominant mode*.

7. Commonly used printed microstrip lines support a hybrid mode, consisting of both TE and TM wave types. However, because the longitudinal components of the electric and magnetic field vectors are small, we can approximately treat the wave as a "quasi-TEM" wave, similar to that in a transmission line.

8. Electromagnetic resonators support oscillating electromagnetic fields. At high frequencies, two types of such resonators are mostly used, the coaxial-line (or

other similar line) resonator and the cavity-type resonator. The latter is obtained as a special case of a waveguide, short-circuited at two ends.

9. The quality factor of waveguide resonators may be by two orders of magnitude greater than that of lumped-element resonant circuits.

# QUESTIONS

**Q23.1.** Write the instantaneous value of $\mathbf{E}_{tot}(x, y, z) = \mathbf{E}(x, y)e^{-\gamma z}$, where $\gamma = \alpha + j\beta$.

**Q23.2.** Complete the derivation of Eq. (23.7).

**Q23.3.** Define in your own words the TEM, TE and TM waves. What does "mode" mean?

**Q23.4.** Can the complex propagation coefficient $\gamma$ in Eq. (23.5) be real? Can it have a real part?

**Q23.5.** What are eigenvalues (characteristic values) of a parameter in a boundary-value problem? What do they depend on?

**Q23.6.** The wave impedance of a TEM wave is always real. Are the wave impedances of TE and TM waves also always real? Explain.

**Q23.7.** Under which conditions is the relation (23.13) valid?

**Q23.8.** What is the physical meaning of the coefficients $m$ and $n$ in the field components inside a rectangular waveguide in Eqs. (23.16) to (23.20)?

**Q23.9.** What is the phase and group velocity in a rectangular waveguide in these three cases? (1) $f < f_c$, (2) $f = f_c$, and (3) $f > f_c$

**Q23.10.** What is the attenuation constant in a rectangular waveguide in these three cases? (1) $f < f_c$, (2) $f = f_c$, and (3) $f > f_c$

**Q23.11.** What are the parameters that determine the cutoff frequency in a waveguide?

**Q23.12.** A signal consisting of frequencies in the vicinity of a frequency $f_1$, and a signal consisting of frequencies in the vicinity of a frequency $f_2$, propagate unattenuated along a rectangular waveguide in the $TE_{10}$ mode. If $f_1 < f_2$, which is faster?

**Q23.13.** What will eventually happen with the signals from the preceding question if the waveguide is long?

**Q23.14.** A signal consisting of frequencies in the vicinity of a frequency $f_1$ propagates along a rectangular waveguide as a $TE_{10}$ mode. What happens if the bandwidth of the signal is relatively large?

**Q23.15.** What are *propagating modes* and *evanescent modes* in a waveguide?

**Q23.16.** You would like to have openings for airing a shielded room (a Faraday's cage) without enabling electromagnetic energy to enter or leave the cavity. You are aware that a field of a certain microwave frequency is particularly pronounced around the room, but you do not know its polarization. Can you make the openings in the form of waveguide sections? What profile of the waveguide would you use?

**Q23.17.** You are using a square waveguide that is bent and twisted along its way. The waveguide is excited with the $TE_{10}$ mode (the $E$ field parallel to the $y$ axis). Can you be certain about the polarization of the wave at the receiving point? Explain.

**Q23.18.** What is the physical meaning of the *dominant mode* in a waveguide?

**Q23.19.** A rectangular waveguide along which waves of many frequencies and modes propagate is terminated in a large metal box. Can you extract from the box a signal of a specific frequency and a desired mode by connecting a section of the same waveguide at another point of the box?

**Q23.20.** How would you construct a high-pass filter (i.e., a filter transmitting only frequencies above a certain frequency), using sections of rectangular waveguides?

**Q23.21.** Propose a method for exciting the $TE_{11}$ mode in a rectangular waveguide.

**Q23.22.** You would like for a rectangular waveguide with a $TE_{10}$ wave to *radiate* (leak) from a series of narrow slots you made in its walls. For this, you need slots that would force the internal waveguide currents to appear on its outer surface. How do the slots need to be oriented to accomplish this?

**Q23.23.** Sketch the electric and magnetic field lines for two microstrip lines, one with a substrate twice the thickness of the other, but with the same permittivity. In which case is the quasi-TEM approximation more accurate? Explain.

**Q23.24.** Sketch the electric and magnetic field lines for two microstrip lines on substrates of equal thicknesses, but where one has a permittivity two times higher than the other. In which case is the quasi-TEM approximation more accurate? Explain.

**Q23.25.** In the resonant circuit of Example 23.10, explain why it is hard to achieve a large $Q$ factor. Why do losses go up as the frequency increases?

**Q23.26.** You would like to have a coaxial-line resonator with as large a $Q$ factor as possible for a given outer resonator size. What would you do?

**Q23.27.** Propose two methods for the excitation and energy extraction from a coaxial resonator.

**Q23.28.** Find the energy contained in coaxial resonators of lengths $\lambda$, $\frac{3}{2}\lambda$, and $2\lambda$, using Eq. (23.36). What is the $Q$ factor of these resonators?

**Q23.29.** Sketch the current and voltage along an open-ended microstrip line resonator that is half of a guided wavelength long. What is the impedance at the center of the resonator, and what at the two ends?

**Q23.30.** What loss mechanisms can you think of in an open-ended microstrip line resonator?

**Q23.31.** A rectangular waveguide with a $TE_{10}$ mode is terminated in a large rectangular cavity (e.g., of a microwave oven). Describe qualitatively what happens.

**Q23.32.** Propose two methods for the excitation and energy extraction from a cavity resonator with a $TE_{101}$ wave type in it.

## PROBLEMS

**P23.1.** Prove that for any TEM wave the electric and magnetic field vectors are normal to each other at all points.

**P23.2.** Prove that at any cross section of a two-conductor transmission line with a forward traveling wave, the ratio of the voltage between the conductors and the current in them equals $Z_{TEM}$ in Eq. (23.7). Show that for two-conductor transmission lines, $C'L' = \epsilon\mu$.

**P23.3.** Prove that since at any cross section of a multiconductor transmission line $\sum Q'(z) = 0$, it follows that $\sum I(z) = 0$, where the sum refers to all the conductors of the line.

**P23.4.** Prove that Eqs. (23.1) to (23.4) imply that the electric and magnetic field vectors of a TE wave are normal to each other at all points.

**P23.5.** Write the instantaneous values of all the components of the $TE_{10}$ wave in a rectangular waveguide. From these equations, sketch the distribution of the E-field and the H-field in the waveguide at $t = 0$.

**P23.6.** Determine the cutoff frequencies of an air-filled waveguide with $a = 2.5$ cm and $b = 1.25$ cm, for the following wave types: (1) $TE_{01}$, (2) $TE_{10}$, (3) $TE_{11}$, (4) $TE_{21}$, (5) $TE_{12}$, and (6) $TE_{22}$.

**P23.7.** Plot the mode impedances between 8 and 12 GHz for an air-filled rectangular waveguide with $a = 2.5$ cm and $b = 1.25$ cm, for the following wave types: (1) $TE_{01}$, (2) $TE_{10}$, (3) $TE_{11}$, (4) $TE_{21}$, (5) $TE_{12}$, and (6) $TE_{22}$.

**P23.8.** Plot the wavelength $\lambda_z$ along a rectangular waveguide with $a = 2$ cm, $b = 1$ cm, and air as the dielectric, if the wave is of the $TE_{10}$ type, for frequencies between 8 and 10 GHz. Is the wavelength shorter or longer than in an air-filled coaxial line?

**P23.9.** Plot the phase and group velocities in problem P23.8.

**P23.10.** In a rectangular waveguide from problem P23.8, two signals are launched at the same instant. The frequency range of the first is in the vicinity of $f_1 = 10$ GHz, and of the second in the vicinity of $f_2 = 12$ GHz. Both signals propagate as $TE_{10}$ waves. Find the time intervals the two signals need to cover a distance $L = 10$ m, and the difference between the two intervals. Which signal is faster?

**P23.11.** Derive Eq. (23.33) starting from the tangential electric field boundary condition.

**P23.12.** The *effective dielectric constant* of a microstrip line depends on its dimensions approximately as

$$\epsilon_e = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \frac{1}{\sqrt{1 + 12\,h/w}},$$

where the parameters are explained in Fig. P23.12. Plot the effective dielectric constant for $h/w$ ratios between 0.1 and 10 (this is the approximate range for practical use), and for substrates that have relative permittivities of 2.2 (Teflon-based Duroid), 4.6 (FR4 laminate), 9 (aluminum nitride), 12 (high-resistivity silicon), and 13 (gallium arsenide).
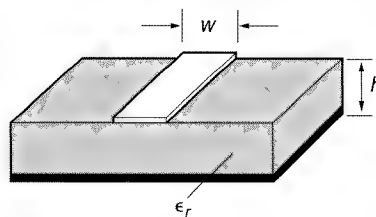


**Figure P23.12** A microstrip line

**P23.13.** The approximate formulas for microstrip line impedance and propagation constant based on the quasi-TEM approximation are given by

$$\beta = \omega\sqrt{\epsilon_0\mu_0}\sqrt{\epsilon_e},$$

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\epsilon_e}} \ln\left(\dfrac{8h}{w} + \dfrac{w}{4h}\right), & \dfrac{w}{h} \leq 1 \\[4mm] \dfrac{120\pi}{\sqrt{\epsilon_e}\{(w/h) + 1.393 + 0.667\ln[(w/h) + 1.444]\}}, & \dfrac{w}{h} > 1 \end{cases}$$

Plot the characteristic impedance as a function of the ratio $w/h$ (between 0.1 and 10), and for the relative permittivities from problem P23.12. What can you conclude about the impedance as the line gets narrower?

**P23.14.** Plot the current, voltage, and impedance along a half-wavelength coaxial resonator short-circuited at both ends. If you want to feed the resonator with another piece of the same kind of cable, at which place along the resonator would you do it and why?

**P23.15.** Plot the current, voltage, and impedance along a half-wavelength coaxial line resonator open-circuited at both ends. You want to feed the resonator with a 50-$\Omega$ coaxial line. Propose (sketch) a way to do it, and explain.

**P23.16.** Determine the maximum possible energy stored in a cubical resonant air-filled cavity with $a = b = d = 10$ cm, at a resonant frequency corresponding to the $TE_{101}$ wave. The electric strength of air is 30 kV/cm.

# 24

# *Fundamentals of Electromagnetic Wave Radiation and Antennas*

## 24.1 Introduction

We know that plane electromagnetic waves propagate through space and carry energy, but we do not still know how such waves can be produced. For the creation of electromagnetic waves we need specific structures with time-varying charges and currents. The process of producing electromagnetic waves, which then propagate with no connection to the sources, is known as *electromagnetic radiation*.

Theoretically, any system containing time-varying charges and currents radiates a certain amount of energy. In many cases in actual practice the radiation can either be ignored or is intentionally suppressed. For example, radiation from 60-Hz (or 50-Hz) power transmission lines exists in theory, but the power radiated is so small that it can practically not be detected. At high frequencies, coaxial cables or hollow waveguides are used to transmit energy precisely because they do not radiate at any frequency.

Specific structures aimed at efficient radiation of electromagnetic waves are referred to as *transmitting antennas*. Typically, transmitting antennas do not radiate

equally in all directions, i.e., they have certain *directional radiation properties*. An entire complicated science of designing and analyzing antennas has been developed in the last hundred years. Although a great number of antennas are available today, the analysis and clever design of antennas for ever-increasing numbers of new applications is an engineering challenge. The present-day powerful numerical methods enable efficient design of many classes of antennas, but a profound knowledge of electromagnetic field theory is needed to make optimal use of such methods.

The electromagnetic energy radiated by an antenna invariably carries a certain signal (information) to be transmitted to one or many receivers. How is energy, and the signal, extracted from an electromagnetic wave? Basically, structures the same as transmitting antennas are used for that purpose, in which case they are called *receiving antennas*. We will see that the most important properties of a receiving antenna can be evaluated if they are known for the same antenna when it is transmitting. In fact, frequently one antenna is used for both transmitting and receiving (e.g., antennas in mobile phones).

Transmitting and receiving antennas are the vocal cords and ears of all radio wave communication systems. Although quite different in shape and size, rabbit-ear antennas and rod antennas for our portable radios, various antennas for TV reception, antennas for the reception of satellite TV programs, antennas used for communication with satellites surveying distant planets, and antennas for astrophysical research that can be the size of an entire valley all operate on the same principles. This chapter is devoted to explaining basic principles and concepts related to radiation and propagation of electromagnetic waves.

## 24.2  Transmitting and Receiving Antennas

Antennas can be used to transmit or receive signals in the form of electromagnetic waves. In either case, the antenna is connected through its feed to some circuit. For example, in reception the antenna is usually followed by a low-noise amplifier because the signals are often very small and the amplifier serves to overcome the noise that the signal is buried in. In transmission, a power amplifier is often connected before the antenna feed. An engineer needs to know how the antenna behaves as part of the rest of the circuit. In the following sections, we discuss transmitting and receiving antennas as circuit elements.

### 24.2.1  NOTES ON TRANSMITTING ANTENNAS

A transmitting antenna takes energy from a source (e.g., an oscillator) and radiates a part of this energy in the form of a free electromagnetic wave. Frequently, the source is connected to the antenna via a transmission line, called the *antenna feed*. Looking from the source (or the feed end terminals), a transmitting antenna is just a receiver of energy. Most frequently, the source feeds the antenna with a time-harmonic voltage, so that in the complex (phasor) domain, the source sees the transmitting antenna as a complex impedance $Z_A$. This impedance is known as the (transmitting) *antenna impedance* and it is, in general, a function of frequency.
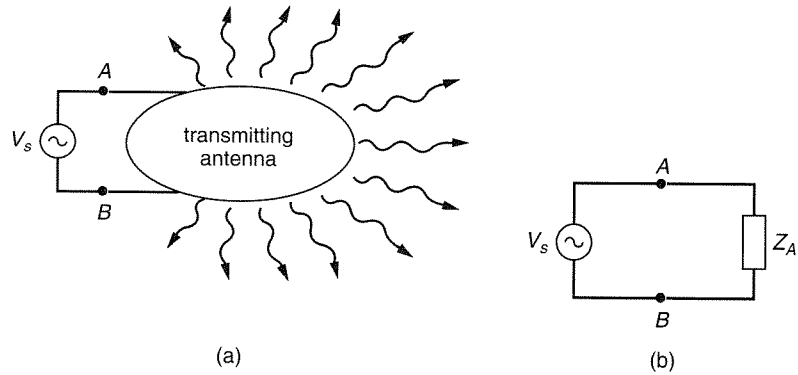
**Figure 24.1** (a) A symbolic picture of a transmitting antenna, and (b) its equivalent circuit as seen by the generator. Wavy arrows of unequal lengths symbolize that the radiation differs in different directions in space.

Neglecting possible losses in the antenna, the time-average power delivered to the antenna is the radiated power. Therefore, the time-average radiated power is given by $P_{\text{rad}} = R_{\text{rad}} I_0^2$, where $R_{\text{rad}}$ is the real part of the antenna impedance, and $I_0$ is the rms value of the current at the antenna terminals.

Most frequently, the transmitting antenna is designed to radiate electromagnetic waves in specific directions, depending on the application. For example, a broadcasting antenna should radiate in all horizontal directions, whereas a satellite antenna should radiate in a very narrow beam toward the corresponding satellite. A symbolic picture of a transmitting antenna and its equivalent circuit are shown in Fig. 24.1.

## 24.2.2 NOTES ON RECEIVING ANTENNAS

A receiving antenna transforms a part of the energy carried by an electromagnetic wave into voltage between two antenna terminals, which are connected to a receiver. This voltage is next amplified and processed as needed. So, from a viewpoint at the receiver terminals, a receiving antenna acts as a voltage generator. In the frequency domain and in complex notation, it behaves as a generator of an emf and an internal impedance, so it can be described by a Thévenin equivalent.

We know that the emf of a Thévenin generator is found as the open-circuit voltage across its terminals (in this case, across the antenna terminals). The internal impedance of the Thévenin generator is that seen by a source connected to the antenna terminals, if the emf (i.e., the incident wave) is not present. This is precisely the *transmitting antenna impedance* mentioned earlier. We reach a very important conclusion:

> *The internal impedance of the Thévenin generator of a receiving antenna is the same as the impedance of the antenna when transmitting.* (24.1)
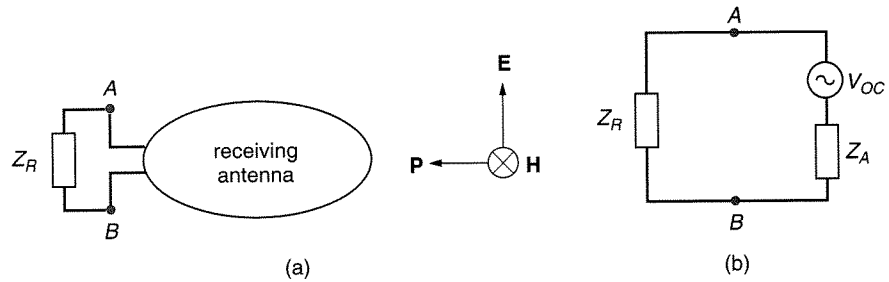
Figure 24.2 (a) A symbolic picture of a receiving antenna, and (b) its equivalent circuit with respect to the receiver. The emf of the equivalent generator depends on the incident wave direction and polarization.

A symbolic picture of a receiving antenna and its Thévenin equivalent with respect to the receiver are sketched in Fig. 24.2.

The Thévenin emf of a receiving antenna depends on the antenna shape and the direction of the incident electromagnetic wave exciting it. For example, assume the receiving antenna to be in the form of two straight wire segments connected to the two receiver terminals (this is known as a *wire dipole antenna*). If the electric field of the wave is parallel to the wire, it is natural to expect the largest emf, and if it is perpendicular, we would expect no emf at all. Thus, although the impedance of the Thévenin generator equivalent to a receiving antenna depends only on the antenna itself, its emf depends greatly on the direction and polarization of the incident electromagnetic wave. So a receiving antenna *has directional properties*, as does a transmitting antenna. We will show that the directional properties of any receiving antenna are known if they are known for the same antenna in transmitting mode.

### 24.2.3   PRINCIPAL ISSUES IN ANTENNA ANALYSIS AND DESIGN

From the preceding simple reasoning, we see that antenna engineers need to know the circuit and radiation properties of antennas in the *transmitting* mode only. In addition to choosing an antenna that radiates properly, they will usually need to match the antenna to the transmitter or receiver, just as we match any passive or active element in circuit theory. Finally, to design a radio communication link, engineers need to know basic properties of electromagnetic-wave propagation in realistic circumstances.

### 24.2.4   ANTENNAS ABOVE CONDUCTING SURFACES

Many antennas are close to approximately flat conducting surfaces, such as antennas above ground, or small antennas on aircraft fuselage or wings, or on cars. If the conducting surface is approximated by a perfectly conducting plane, the antenna can be analyzed by image theory. Figure 24.3 shows a few typical cases. The images at corresponding points should have opposite charges, and currents in opposite directions,

(a)                              (b)                              (c)
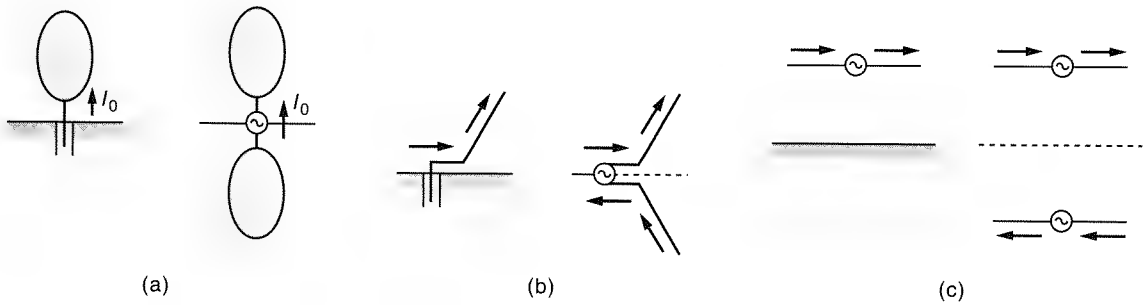
**Figure 24.3** Three examples of antennas close to a perfectly conducting plane and their equivalents with respect to the upper half-space, showing images with the necessary charges and currents: (a, b) monopole antennas fed at the ground plane; (c) a dipole antenna above a ground plane

with respect to the conducting plane. Such charges and currents guarantee a zero tangential component of the electric field vector on the perfectly conducting plane, i.e., the original boundary conditions remain satisfied.

Antennas of the type (c) in Fig. 24.3 are known as *dipole antennas* (antennas with two poles, i.e., two arms). Antennas of the types (a) and (b), which are excited at the surface as indicated, have only one arm, and are called *monopole antennas*. Note that in reality, the radiation field exists *only in the upper half-space*. This means that for cases (a) and (b) in Fig. 24.3, the voltages driving the antennas with their images have to be twice those of the original monopole antennas. Since the current is the same, this means that the impedance of a monopole antenna above a perfect ground equals half the impedance of the corresponding dipole antenna. Note that this conclusion does not apply to the dipole antenna (c), although its impedance will also be different from that of the isolated antenna, due to the presence of the conducting plane.

*Questions and problems:*    Q24.1 to Q24.3, P24.1

## 24.3    Electric Dipole Antenna (Hertzian Dipole)

The electric dipole antenna, or the Hertzian dipole, is the simplest of all radiating systems. It consists of a straight, thin wire conductor of length $l$ with two small conducting spheres or disks at the ends, as sketched in Fig. 24.4. The spheres serve as capacitor electrodes, and make the current $i(t)$ along the dipole wire constant along its length. A sinusoidal generator of angular frequency $\omega$ is connected somewhere along the wire. The derivation of the electric and magnetic field vector components is given in most higher-level electromagnetics textbooks (see, e.g., S. Ramo et al., *Fields and waves in communication electronics*, 3d ed., section 12.3, John Wiley & Sons, 1993). For our introduction to Hertzian dipoles, it is sufficient to quote the expressions so that we can make some important conclusions about the radiated fields. Note that the dipole radiates in a sphere around it, and therefore it is natural to use the spherical coordinate system to describe the radiated fields. At distances far from the dipole in terms of the free-space wavelength (more than about 10 wavelengths), and assuming
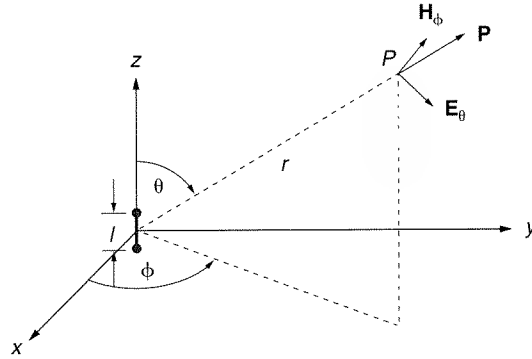
Figure 24.4 The electric dipole antenna (the Hertzian dipole)

that the dipole is situated in a homogeneous dielectric of parameters $\epsilon$ and $\mu$, the complex expressions for the electric and magnetic fields are given by

$$
E_\theta(r, \theta) = \frac{j\beta Il \sin\theta}{4\pi r} \sqrt{\frac{\mu}{\epsilon}} e^{-j\beta r}, \tag{24.2}
$$

$$
H_\phi(r, \theta) = \frac{j\beta Il \sin\theta}{4\pi r} e^{-j\beta r}. \tag{24.3}
$$

*(Electric and magnetic fields far from a Hertzian dipole)*

Thus, the only component of the electric field is in the direction of the unit vector $\mathbf{u}_\theta$, and that of the magnetic field in the direction of the unit vector $\mathbf{u}_\phi$. (In Fig. 24.4, at point $P$ the latter is directed into the paper.) All the other components are zero. Hence the ratio of the amplitudes of the two vectors is the same as for a plane wave,

$$
\frac{E_\theta(r, \theta)}{H_\phi(r, \theta)} = \eta = \sqrt{\frac{\mu}{\epsilon}}. \tag{24.4}
$$

*(Relation between E and H fields of the wave radiated by a Hertzian dipole)*

This was to be expected because at large distances from the dipole the spherical wave is locally a plane wave. In addition, we note that the vectors $E_\theta$ and $H_\phi$ are normal to each other, and that the Poynting vector is directed away from the dipole, as expected, because the radiated wave carries power away from the antenna.

Note that both field components depend on the distance $r$ from the dipole as $1/r$. No static field has this dependence on $r$. This type of field is thus different from any electromagnetic field we considered so far, and is termed the *radiation field*, or the *far field*, of the Hertzian dipole. We are interested here only in this field (the field
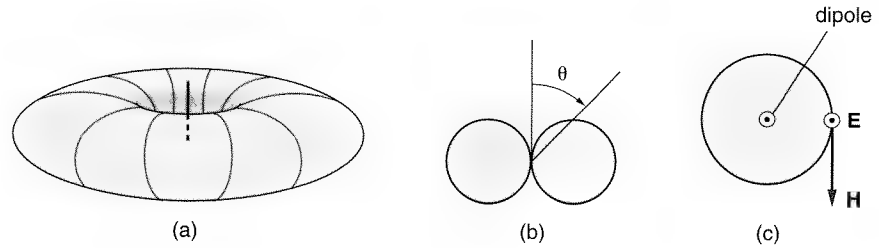
Figure 24.5 Radiation patterns of a Hertzian dipole versus angle $\theta$, normalized to the maximal radiation. (a) The three-dimensional pattern, (b) the $E$-plane pattern, and (c) the $H$-plane pattern are plotted.

closer to the dipole has other components in addition). Because all antennas can be considered as large assemblies of Hertzian dipoles, *the far fields of all antennas also depend on r as 1/r.*

**Example 24.1—Spatial distribution of radiation from a Hertzian dipole.** Note that the simplest radiating element, the Hertzian dipole, does not radiate equally in all directions. This is evident from the factor $\sin\theta$, which tells us that the radiation is the strongest in the plane of symmetry of the dipole (for $\theta = \pi/2$), and zero along the direction defined by the dipole wire ($\theta = 0$ and $\theta = \pi$). Consequently, no real antenna can radiate equally in all directions. To characterize the distribution of radiated field in different directions, it is customary to normalize the field with respect to its maximal value, and to plot a graph of this function. Such a graph is known as the *antenna radiation pattern*. Antenna radiation patterns may be plotted in three dimensions, as in Fig. 24.5a, but it is usually more convenient to plot cuts of the three-dimensional pattern. Usual cuts are those containing the **E** vector (known as the *E-plane pattern*, Fig. 24.5b) or the **H** vector (the *H-plane pattern*, Fig. 24.5c). We will see that with appropriate shapes of antennas, we can obtain a great variety of radiation patterns.

Any radiating system can always be subdivided into a large number of Hertzian dipoles, provided that the current distribution of the system is known. It is important to understand that if this simplest radiating system produces a radiation field with the electric and magnetic field as in a plane wave propagating along the (local) $r$ axis, by superposition (and because of linearity) the same will be true for *any* radiating structure.

**Example 24.2—The half-wave dipole antenna.** A frequently used antenna is in the form of a straight wire dipole of total length equal to about half a wavelength. For such a dipole, it turns out that the current distribution can be roughly approximated by a sine function (Fig. 24.6a). At an introductory level, it is important to remember that the impedance of a half-wave dipole is, very roughly, $73\,\Omega$. The radiation pattern of a half-wave dipole is very similar to that of the Hertzian dipole. The only difference is that the $E$-plane pattern is not in the form of a "number eight" consisting of two circles, but instead of an "eight" consisting approximately of two ellipses, as in Fig. 24.6b.

*Questions and problems:* Q24.4 to Q24.8, P24.2 to P24.5
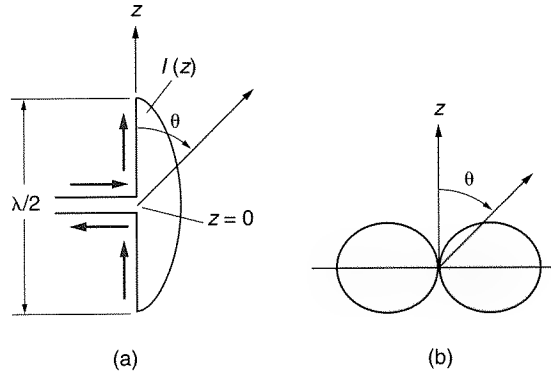
Figure 24.6 (a) The half-wave dipole antenna;
(b) $E$-plane radiation pattern of the dipole

## 24.4   Antenna Directivity

The radiation pattern is only one of several parameters that describe the spacial distribution of antenna radiation. We now describe another quantity, which is used more frequently than radiation patterns.

Assume that an antenna radiates equally in all directions. Such a hypothetical antenna is known as an *isotropic*, or *omnidirectional, antenna*. For an isotropic antenna, the power density is the same for all points of a sphere of radius $r$ centered at the antenna. This power density is denoted by $S_i$, and is equal to

$$S_i = \frac{P_{rad}}{4\pi r^2},$$

in watts per m². We can now define the antenna *directivity* as the ratio of the power density radiated in a given direction $(\theta, \phi)$ to the isotropic power density:

$$D(\theta, \phi) = \frac{S(\theta, \phi)}{S_i} = \frac{4\pi r^2 S(\theta, \phi)}{P_{rad}}, \qquad P_{rad} = R_{rad}|I_0|^2. \tag{24.5a}$$

*(Definition of directivity of an antenna)*

From this expression, we see that the directivity of an isotropic antenna is unity.

We can relate the directivity to the electric and magnetic fields through the Poynting vector. From the Poynting theorem, Eq. (19.42), the surface integral of the Poynting vector is the radiated power. The power density is therefore equal to the magnitude of the Poynting vector, which we remember is $\mathbf{E} \times \mathbf{H}$. We know that the electric and magnetic far fields at a point are proportional to $1/r$, where $r$ is the distance of the observation point from the antenna. Consequently, the time-average Poynting vector, $\mathcal{P}(r, \theta, \phi)$, at a point in the far field is proportional to $1/r^2$, and it is also proportional to the power radiated by the antenna, $P_{rad}$. So we get an alternate

expression for the antenna directivity *relative to an isotropic antenna*:

$$D(\theta, \phi) = \frac{4\pi r^2 |\mathcal{P}(r, \theta, \phi)|}{P_{\text{rad}}} = \sqrt{\frac{\epsilon}{\mu}} \frac{4\pi r^2 |\mathbf{E}(r, \theta, \phi)|^2}{P_{\text{rad}}}, \qquad P_{\text{rad}} = R_{\text{rad}} |I_0|^2. \quad (24.5b)$$

*(Definition of directivity of an antenna)*

We see that the directivity *depends only on spherical angles $\theta$ and $\phi$* and can be used as a measure of the antenna directional properties. Often, the directivity is given in decibels,

$$[D(\theta, \phi)]_{\text{dB}} = 10 \log\{D(\theta, \phi)\} \quad (\text{dB}). \tag{24.6}$$

*(Directivity in decibels)*

The multiplier of the logarithm is 10, not 20, because the directivity is defined through power, not field intensity.

The plot of the directivity in space is known as the *antenna power pattern*. As in the case of field patterns, more frequently the antenna *E*-plane or *H*-plane power pattern or both are plotted, although the patterns in other planes may also be of interest. From the definition in Eqs. (24.5), we see that the power pattern is proportional to the *square* of the field pattern.

If, in referring to the directivity, the direction (defined by angles $\theta$ and $\phi$ or in some other way) is not specified, by convention this means that *the maximum value of the directivity is implied*, i.e.,

$$D = [D(\theta, \phi)]_{\text{max}}. \tag{24.7}$$

*(Definition of directivity with no reference to direction)*

In this text, we will use the term "maximal directivity" for clarity. It is of significant practical interest because in many applications antennas are used to radiate in a specific direction (for example, in satellite links).

Manufacturers often specify *antenna gain* and maximum antenna gain, $G(\theta, \phi)$ and $G = [G(\theta, \phi)]_{\text{max}}$, instead of directivity. The difference between these two parameters is that the antenna gain includes any power losses in the antenna (such as losses due to the finite conductivity of the metal that the antenna is made of). Therefore, the gain of an antenna is a number that is at most as large as the directivity of the same antenna.

**Example 24.3—Directivity of the Hertzian dipole.** From the far electric field of the Hertzian dipole, the power radiated by the dipole, and hence its directivity, can be calculated using Poynting's theorem. The derivation is given in most higher-level electromagnetic textbooks, and here we give only the final expression for the directivity of the Hertzian dipole:

$$[D(\theta, \phi)]_{\text{Hertzian dipole}} = 1.5 \sin^2 \theta. \tag{24.8}$$

*(Directivity of Hertzian dipole)*

This is proportional to the square of the electric field pattern, which is proportional to $\sin\theta$. The maximal directivity of the Hertzian dipole is thus

$$D_{\text{Hertzian dipole}} = 1.5. \tag{24.9}$$

<div align="right"><em>(Maximal directivity of Hertzian dipole)</em></div>

Note that this means that the power density (magnitude of the Poynting vector) of the field radiated by the Hertzian dipole in the plane $\theta = \pi/2$ is simply 1.5 times that of an isotropic antenna radiating the same power and located at the position of the dipole. The maximal radiated electric field strength of the Hertzian dipole is thus $\sqrt{1.5} \simeq 1.22$ times that of an isotropic antenna.

**Example 24.4—Directivity of the half-wave dipole.** To determine the directivity of the half-wave dipole, we need to know the radiation field of the dipole. Because we know the (approximate) current distribution along it, this is not too complicated, but at this introductory level we will skip the derivation. The final result for the directivity is

$$D(\theta) = \sqrt{\frac{\mu}{\epsilon}} \, \frac{1}{\pi R_{\text{rad}}} \, \frac{\cos^2(\pi/2\cos\theta)}{\sin^2\theta}. \tag{24.10}$$

where $\theta$ is the angle between the dipole axis ($z$ axis in Fig. 24.6a), and the direction toward the point in the radiation field. $R_{\text{rad}}$ is the antenna resistance (we know that for a half-wave dipole, $R_{\text{rad}} \simeq 73\,\Omega$). Therefore the maximal directivity of the half-wave dipole is

$$D = D(\pi/2) = \frac{120}{73} \simeq 1.64. \tag{24.11}$$

<div align="right"><em>(Maximal directivity of half-wave dipole)</em></div>

*Questions and problems:*    Q24.9 and Q24.10

## 24.5    The Receiving Antenna

A receiving antenna is always very far from the transmitting antenna (practical reasons for this fact are given in Chapter 25). Therefore, the current induced in it is very small compared to that in the transmitting antenna. The field produced at the transmitting antenna by the current induced in the receiving antenna is evidently negligible. Therefore, the equivalent circuit for the transmitting-receiving antenna system is as in Fig. 24.7.

The transfer impedance (or mutual impedance), $Z_{12}$, has exactly the same meaning as in circuits coupled by the induced electric field ("magnetic coupling"). In this case also the induced electric field is the one that induces the emf and current in the receiving antenna, the only difference being that the finite velocity of propagation of electromagnetic waves must now be taken into account. Note that in circuits the effect of finite wave velocity can normally be neglected because all the parts of a circuit are close when compared with the wavelength.

Although this is not necessary for the following derivations, assume that both antennas are matched to the feed lines, as indicated in Fig. 24.7, which is most often

$$|Z_{12}| \ll |Z_1|, |Z_2|$$
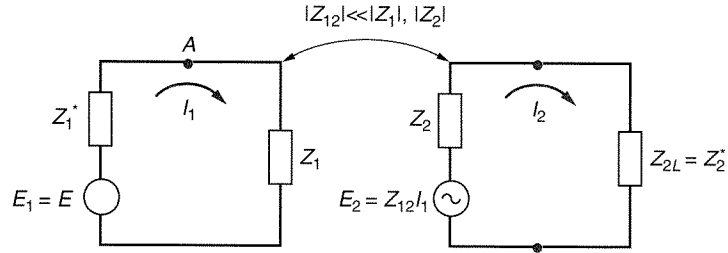


**Figure 24.7** Equivalent circuit for the transmitting-receiving antenna system, when antenna 1 is transmitting

the case. If antenna 1 is transmitting and is excited by a generator of voltage $V$, the current in the receiving antenna, antenna 2, is given by (see Fig. 24.7)

$$I_{2\text{rec}} = \frac{Z_{12}(\theta_1, \phi_1, \theta_2, \phi_2)I_1}{2R_2} = Z_{12}(\theta_1, \phi_1, \theta_2, \phi_2)\frac{V}{4R_2R_1}. \tag{24.12a}$$

The angles $\theta_1$ and $\phi_1$ are local spherical angles of antenna 1, defining the direction from antenna 1 toward antenna 2. Similarly, $\theta_2$ and $\phi_2$ are local spherical angles of antenna 2, defining the direction from antenna 2 toward antenna 1.

If the generator is moved to antenna 2, and antenna 1 is receiving, the current in antenna 1 is similarly

$$I_{1\text{rec}} = Z_{21}(\theta_1, \phi_1, \theta_2, \phi_2)\frac{V}{4R_1R_2}. \tag{24.12b}$$

Note that we first determined the current in branch 2 of the (coupled) circuits due to the generator in branch 1. We next moved the generator to branch 2, and determined the current in branch 1. This is a typical example of circuit reciprocity, which dictates that the currents in Eqs. (24.12a) and (24.12b) must be the same. Consequently, $Z_{21}(\theta_1, \phi_1, \theta_2, \phi_2) = Z_{12}(\theta_1, \phi_1, \theta_2, \phi_2)$.

The transfer impedances implicitly contain the direction of radiation of the transmitting antenna, as well as the direction from which the incident wave is arriving to the receiving antenna. Therefore the transfer impedance contains the directional properties of the two antennas in both the transmitting and the receiving mode. For $Z_{21} = Z_{12}$ to hold in all cases, the transmitting and receiving patterns of an antenna must be the same. We reach an extremely important conclusion:

*The receiving pattern of a receiving antenna is the same as the radiation pattern of the same antenna in transmitting mode.* (24.13)

The antenna's *effective area* is a quantity used frequently for describing a receiving antenna. Assume a receiving antenna 2 *matched to its load*. Assume that the incident wave arrives from the direction defined by the angles $\theta_2$ and $\phi_2$ with respect to

the receiving antenna spherical coordinate system. The power density (time-average of the Poynting vector) at the antenna terminals is $S(\theta, \phi) = \mathcal{P}_1(\theta, \phi) = E_1^2/\eta$, where $E_1$ is the rms value of the electric field vector due to the transmitting antenna 1 in the direction $(\theta, \phi)$. Assume, also, that vector $\mathbf{E}_1$ is oriented so that the emf induced in the receiving antenna is the largest possible. The effective area of the antenna, $A_{\text{eff}}(\theta_2, \phi_2)$, is then defined by the equation

$$
A_{\text{eff2}}(\theta, \phi) = \frac{P_{2 \text{ matched load, optimal reception}}}{S(\theta, \phi)}
$$

$$
= \frac{P_{2 \text{ matched load, optimal reception}}}{\mathcal{P}_1(\theta, \phi)}. \tag{24.14}
$$

*(Definition of the effective area of a receiving antenna)*

*Questions and problems:*   Q24.11

## 24.6   The Friis Transmission Formula

Let us now examine a line-of-sight link (i.e., a radio link in which two antennas can "see" each other) between two antennas, as in Fig. 24.8. A transmitting antenna, of directivity $D_1(\theta_1, \phi_1)$ in the direction of the receiving antenna, radiates a power $P_{1\text{rad}}$, and a receiving antenna, with an effective area $A_2(\theta_2, \phi_2)$ in the direction of the transmitter, is matched to the receiver. The power delivered to the load connected to the receiving antenna terminals, from Eq. (24.14), is

$$
P_{2 \text{ matched load, optimal reception}} = A_{\text{eff2}}(\theta_2, \phi_2)\mathcal{P}_1,
$$

or, using Eq. (24.5*b*),

$$
P_{2 \text{ matched load, optimal reception}} = A_{\text{eff2}}(\theta_2, \phi_2)\frac{D_1(\theta_1, \phi_1)P_{1\text{rad}}}{4\pi r^2},
$$
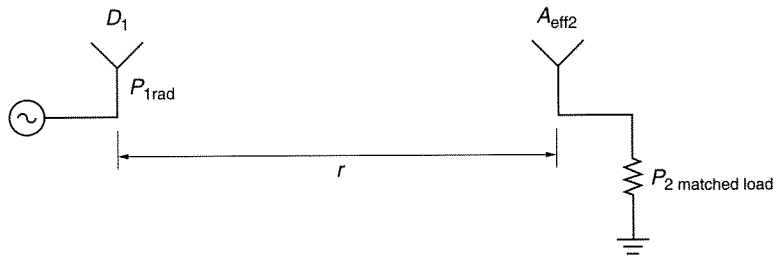


Figure 24.8 A line-of-sight link between two antennas

so that

$$\text{A}_{\text{\tiny EFF2}}$$

$$P_{2\text{ matched load, optimal reception}} = P_{1\text{rad}} \frac{D_1(\theta_1, \phi_1) A_2(\theta_2, \phi_2)}{4\pi r^2}. \tag{24.15}$$

*(Friis transmission formula)*

This is known as the *Friis transmission formula* and it describes the power transmission in a line-of-sight antenna link. Note that we expressed it in terms of the directivity of the transmitting antenna, and the effective area of the receiving antenna.

## The Relation Between Directivity and Effective Area

We can use the Friis transmission formula to show that the effective area of a receiving antenna can be expressed in terms of the antenna directivity, i.e., *in terms of the properties of the antenna in transmitting mode*. In the link shown in Fig. 24.8, we first assume antenna 1 is transmitting, and antenna 2 is receiving. The Friis formula gives for this case (labeled with a prime):

$$P'_{2\text{rec}} = P'_{1\text{rad}} \frac{D_1 A_{\text{eff2}}}{4\pi r^2}, \tag{24.16}$$

where the $(\theta, \phi)$ dependence was omitted for brevity, i.e., the expression is valid for any direction. Now let us look at the second case (labeled with a double prime), when antenna 2 transmits, and antenna 1 receives. The Friis formula is now

$$P''_{1\text{rec}} = P''_{2\text{rad}} \frac{D_2 A_{\text{eff1}}}{4\pi r^2}. \tag{24.17}$$

Now we can apply reciprocity, as discussed in section 24.5. If instead of currents and voltages, we apply the reciprocity to transmitted and received powers, we obtain

$$\frac{P'_{2\text{rec}}}{P''_{1\text{rec}}} = \frac{P'_{1\text{rad}}}{P''_{2\text{rad}}}.$$

[This is obtained, with a little manipulation, from Eqs. (24.12a) and (24.12b); see problem P24.7.] Now Eqs. (24.16) and (24.17) are substituted in the last equation, and after canceling the powers and the term $4\pi r^2$, we obtain an interesting result:

$$\frac{A_{\text{eff1}}(\theta, \phi)}{D_1(\theta, \phi)} = \frac{A_{\text{eff2}}(\theta, \phi)}{D_2(\theta, \phi)}.$$

Here we have inserted the angular dependence again so as not to forget that $D$ and $A_{\text{eff}}$ change when the angle between the two antennas changes. What does this equation tell us? We did not assume anything about the two antennas, and we conclude that the ratio of the effective area and the directivity is always the same constant! If we knew what this constant was, we could always relate the directivity to the effective area, and therefore, the transmitting properties of an antenna to its receiving properties.

It is relatively easy to show that the integral of the directivity over all angles (sphere of any radius) is $4\pi$. It can also be shown (but this is not at all easy and is

well beyond the scope of this book) that the integral of the effective area over all angles is equal to $\lambda^2$, where $\lambda$ is the free-space wavelength. Then the ratios in the last equation are the same as the ratios of their integrals, or $\lambda^2/(4\pi)$, giving

$$A_{\text{eff}}(\theta, \phi) = \frac{\lambda^2}{4\pi} D(\theta, \phi). \tag{24.18}$$

*(Relationship between antenna effective area and directivity)*

This equation tells us that the directivity of an antenna is proportional to the effective area (and therefore size) of the antenna measured in free-space wavelengths.

**Example 24.5—Effective area of a half-wave dipole with sinusoidal current distribution.** From Eq. (24.10) in Example 24.4, we know the directivity of the half-wave dipole. From Eq. (24.18), the effective area of the half-wave dipole is thus

$$[A_{\text{eff}}(\theta)]_{\text{half-wave dipole}} = \sqrt{\frac{\mu}{\epsilon}} \frac{\lambda^2}{4\pi^2 R_{\text{rad}}} \frac{\cos^2(\pi/2 \cos\theta)}{\sin^2\theta}.$$

The maximal possible power delivered to the antenna is obtained if the wave is incident from the direction $\theta = \pi/2$ (and if, of course, the electric field vector is parallel to the dipole). In that case, with $R_{\text{rad}} = 73\,\Omega$, the preceding expression yields

$$[A_{\text{eff}}(\pi/2)]_{\text{max half-wave dipole}} \simeq 0.13\,\lambda^2.$$

This means that a matched half-wave dipole extracts from the wave the power contained in approximately $\lambda^2/8$ of the wavefront, as sketched in Fig. 24.9.

**Example 24.6—Directivity of a 3-meter dish antenna.** As another example, consider a 3-meter diameter dish for satellite TV at 12 GHz. Assuming the effective area of the dish (for $\theta = 0°$) is the same as its geometric area (which is approximately true for reflector antennas),
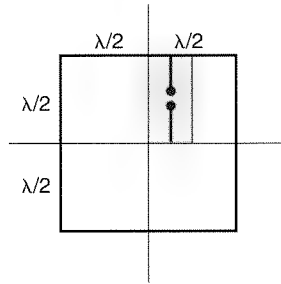


Figure 24.9 A matched half-wave dipole extracts from the wave the power contained in approximately $\lambda^2/8$ of the wavefront (shaded area with the sketch of the dipole)

from Eq. (24.17) we obtain $D = (4\pi \cdot \pi \cdot 1.5^2)/(0.025^2) = 141,975$. From this example, we see that antennas that are large when measured in wavelengths have high directivities, i.e., narrow beams.

**Example 24.7—A radio communication link with half-wave dipole antennas.** Assume that both the transmitting and receiving antennas in a line-of-sight link are half-wave dipoles a distance $r$ apart. Let the dipoles be parallel to each other, and normal to the line connecting them, which ensures the maximal possible transmission of power under matched conditions. Specifically, let the frequency be $f = 900\,\text{MHz}$ (a cellular phone frequency), and the distance between the antennas $r = 100\,\text{m}$. The wavelength corresponding to $f = 900\,\text{MHz}$ is $\lambda = c/f = 0.333\,\text{m}$. We can express the Friis transmission formulas only in terms of the directivities, using Eq. (24.18):

$$P_{2\text{ matched load, optimal reception}} = P_{1\text{rad}}\frac{D_1 D_2}{r^2\lambda^2}. \tag{24.19}$$

The ratio of the power received by the receiver matched load and the power radiated by the transmitting antenna is in this case

$$\frac{P_{2\text{ matched load, optimal reception}}}{P_{1\text{rad}}} = \left(\frac{D}{r\lambda}\right)^2 = \left(\frac{0.333}{4\pi\ 100}\right)^2 1.64^2 \simeq 1.9\cdot 10^{-7}.$$

Thus, for a transmitted power of 10 W, the received power will be only about $1.9\,\mu\text{W}$.

**Example 24.8—Other forms of the Friis formula.** The Friis formula can also be expressed in terms of effective areas. From Eqs. (24.15) and (24.18) we obtain

$$P_{2\text{ matched load, optimal reception}} = \frac{A_{\text{eff1}}(\theta_1, \phi_1)A_{\text{eff2}}(\theta_2, \phi_2)}{(\lambda r)^2}\ P_{1\text{rad}}. \tag{24.20}$$

Usually, the quantity given for an antenna by the manufacturer is the maximal directivity in decibels, $D_{\text{dB}}(\theta, \phi) = 10 \log D(\theta, \phi)$. Also, commonly the radiated and received power are expressed with respect to a certain reference power level, the most frequent being 1 mW or 1 W. As an example, let the reference power level be 1 mW. We divide Eq. (24.19) by 1 mW, take the decimal logarithm of both sides, and multiply by 10 (to obtain decibels as calculated for power). The Friis formula expressed in decibels thus becomes

$$P_{2\text{ matched load, optimal reception, dBm}} = P_{1\text{rad, dBm}} + D_{1\text{dB}}(\theta_1, \phi_1) + D_{2\text{dB}}(\theta_2, \phi_2) + 20\log\frac{\lambda}{r} - 21.984. \tag{24.21}$$

where the subscript "dBm" stands for "decibels over one milliwatt," and $-21.984 = 20\log(1/4\pi)$. As an example of directivity expressed in decibels, the directivity of the satellite dish from Example 24.6 is $10\log 141,975 = 51.5\,\text{dB}$.

*Questions and problems:*    Q24.12, P24.6 to P24.9

## 24.7    Brief Overview of Other Antenna Types and Additional Concepts

There is a very wide variety of antennas used for different frequency ranges and different purposes. We conclude this brief chapter with a description of some frequently used antenna types. In addition, we will define some of the important antenna parameters that were not mentioned so far.

An antenna is said to be *narrowband* if it can efficiently emit and receive signals of frequencies in a relatively narrow frequency range, not exceeding more than a few percent of the central frequency. Antennas that can emit and receive efficiently in a broader frequency range are known as *broadband* antennas.

Antennas may have radiation patterns with several maxima. The largest is known as the *main lobe*, and the others as the *side lobes*. The *sidelobe level*, usually in decibels, is the level of the sidelobes with respect to the main lobe. The control of sidelobe levels frequently is not very strict, but in some applications it may be quite strict, requiring sidelobe levels of less than, for example, $-40\,\text{dB}$.

An important property of the main antenna beam is the *beamwidth*. By definition, this is the angle between the directions for which the field intensity is $1/\sqrt{2} = 0.707$ that at the beam maximum. This corresponds to half the power density at the beam maximum.

Antennas are given additional descriptions relating principally to the frequency range in which they are used. We encounter low frequency (LF) antennas (30 kHz to 300 kHz), medium frequency (MF) antennas (300 kHz to 3 MHz), high frequency (HF) antennas (3 MHz to 30 MHz), very high frequency (VHF) antennas (30 MHz to 300 MHz), and ultra high frequency (UHF) antennas (300 MHz to 3000 MHz). The term "microwave antennas" usually implies frequencies above about 3000 MHz. Although some antennas can be used in several frequency ranges, these names also point to certain antenna types.

Sketched in Fig. 24.10a and 24.10b are two basic antenna types we have already met: the wire dipole and the wire monopole antenna. They are used in various forms at virtually all frequencies.

The antenna in Fig. 24.10c is known as the *loop antenna*. It does not radiate (and therefore does not receive a signal) normal to its plane, and the received signal is maximal from the directions in that plane. Therefore two or three such antennas at different points can be used for locating a transmitting antenna ("direction finding"). It is used at frequencies below about 1 GHz.

A great variety of relatively simple directional antennas are used for TV reception (6-MHz-wide channels in several frequency bands in the range 54 MHz to 890 MHz). One such narrowband antenna is the Yagi-Uda array sketched in Fig. 24.10d. It consists of a driven dipole backed by a *passive* (not directly excited) wire, known as the *reflector*, and several passive wires in front of the dipole (toward the transmitter), known as the *directors*. The electromagnetic field induces currents in the passive elements to influence the antenna radiation pattern. If properly designed, a Yagi antenna radiates a relatively narrow beam, but it is a frequency-sensitive structure. Until recently, the design of Yagi arrays was mostly experimental, but nowadays computer codes exist that enable the analysis and design of Yagi antennas with great precision.

(a)        (b)              (c)                    (d)              (e)



(f)                    (g)                    (h)                    (i)
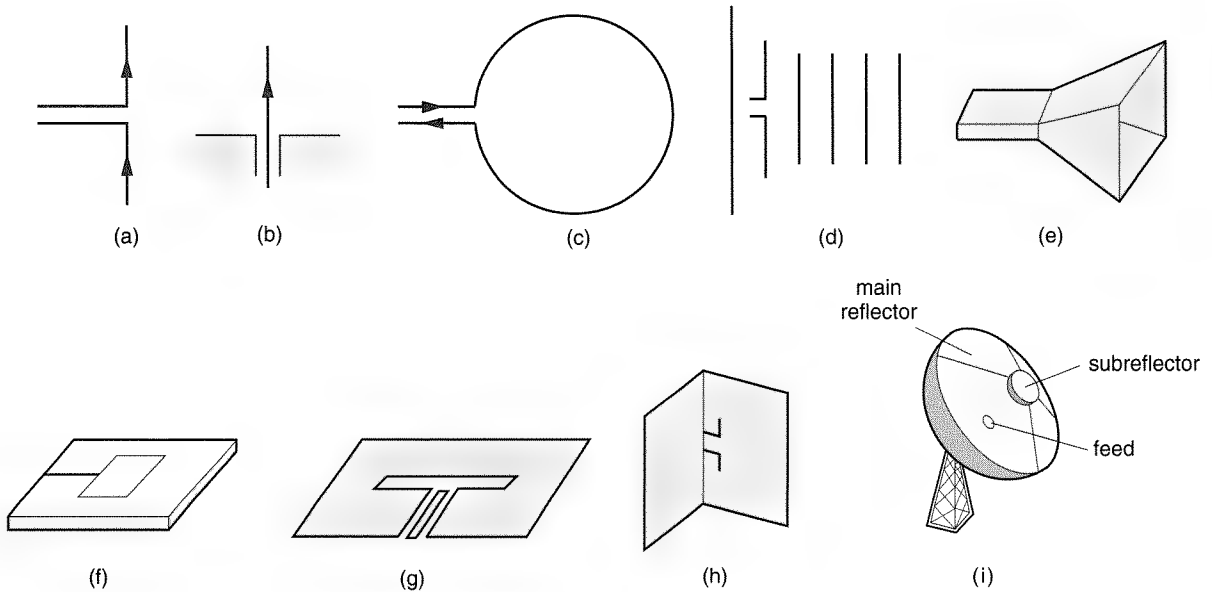
Figure 24.10 Sketches of some basic antennas: (a) a wire dipole; (b) a wire monopole; (c) a loop; (d) a Yagi-Uda array; (e) a horn; (f) a microstrip patch; (g) a slot; (h) a corner reflector; and (i) a parabolic reflector (dish)

If we terminate a rectangular waveguide in a rectangular open horn, as in Fig. 24.10e, the horn enhances the radiation from the open waveguide end. Such antennas are known as *horn antennas*, and are used exclusively as microwave antennas. The radiation of such antennas can be analyzed by assuming a field distribution over the horn aperture, from which, using certain advanced electromagnetic theory methods, the antenna far field can be calculated. Therefore, such antennas are also called *aperture antennas*.

Figure 24.10f is a sketch of the so-called *microstrip patch antenna*. On a dielectric substrate with ground metallization on the other side, a metallic patch is made and fed with a narrow metallic strip. The other terminal of the generator is connected to the ground metallization. (The transmission line obtained in this way is known as a *microstrip line*.) Microstrip patch antennas are narrowband, but due to their flat shape they have many applications where flush mounting is desirable (sometimes these flush-mounted antennas are called "conformal").

An entirely different family of antennas are *slot antennas*, obtained by cutting slots of various shapes in metal screens, as in Fig. 24.10g. There is a theory that enables the analysis of radiation and circuit properties of slot antennas in terms of those of "complementary" antennas, i.e., metallic antennas in the form of the slot, with the screen removed. Slot antennas are used as microwave antennas.

To enhance radiation in a specific direction, metallic reflectors are used as in Fig. 24.10h, where a dipole antenna is backed by a *corner reflector*, and Fig. 24.10i, where a parabolic reflector is used to concentrate the antenna radiation into a very narrow beam, sometimes termed a "pencil beam." Whereas antennas of the form in

Fig. 24.10h are used above about 100 MHz, antennas with a parabolic reflector are meaningless for frequencies below about 1 GHz, because the reflector must be many wavelengths in diameter in order to concentrate the radiation efficiently.

**Example 24.9—High-directivity antennas.** Since $D = (4\pi/\lambda^2)\,A_{\text{eff}}$, antennas with high directivity are large when measured in wavelengths. For example, an antenna that has a maximal directivity of 30 dB should be at least $1000/4\pi \simeq 80$ wavelengths square, or if it were a square, about 9 wavelengths on the side. At 300 MHz, this would be about 9 meters on the side, at 1 GHz about 2.7 m on the side, and at 30 GHz about 9 cm on the side. So, if an airplane or a satellite needs to carry a highly directional antenna, choosing a higher frequency of operation allows for a smaller and lighter antenna, which translates to less fuel, more room, and ultimately lower cost.

When high directivity is required, antenna arrays are frequently used instead of a single very large antenna. In antenna arrays, a large number of smaller, low-directivity elements are fed with a common feed, which makes the array look like one antenna. The elements are usually about a half wavelength apart. This antenna has a geometric area that can be many wavelengths across, and can therefore have a very high directivity. Further, the radiation pattern can be tailored to satisfy different requirements at different angles. For example, an array in one dimension (say, a 10 by 1 array) of antennas will have a high directivity in the plane of the 10 elements, and a low directivity in the orthogonal plane, and its effective area will be very roughly $5\lambda^2$.

## 24.8 Chapter Summary

1. Antennas are metallic and/or dielectric structures used for radiation and reception of electromagnetic waves. As a rule, they have two closely spaced terminals in the circuit-theory sense.

2. A transmitting antenna excited by a sinusoidal generator behaves as a load of a certain frequency-dependent impedance, known as the *antenna impedance*.

3. The simplest antenna is the short electric dipole antenna, or the Hertzian dipole. The current distribution along the Hertzian dipole is assumed to be constant. The radiation properties of the Hertizan dipole can be obtained in a relatively simple manner.

4. Antennas do not radiate equally in all directions. The directional radiation properties of antennas are described by two basic quantities, the antenna radiation pattern and its directivity. The maximal directivity is usually of particular interest.

5. The gain of an antenna is at most as high as its directivity, since it includes any losses in the antenna.

6. If an antenna is used for receiving, it transfers a part of the energy of the incident electromagnetic wave to its load (which is usually the input to an amplifier). A receiving antenna behaves with respect to the load as an equivalent Thévenin generator and has the same internal impedance as when it is transmitting.

7. Directional properties of a receiving antenna are the same as those when it is used for transmitting.

## QUESTIONS

**Q24.1.** You have a black box with two terminals. You connect a generator to these terminals and find out that the black box behaves as an impedance. Can you check by observing the measured impedance whether the terminals belong to a transmitting antenna inside the box? Explain.

**Q24.2.** You have a black box with two terminals. You connect a load to these terminals and find out that the black box behaves as a generator. Can you check by observing the measured current in the load whether the terminals belong to a receiving antenna inside the box? Explain.

**Q24.3.** Why are the images of antennas in Fig. 24.3 as indicated?

**Q24.4.** On many short antennas there are small conducting balls at each end. What are these balls for?

**Q24.5.** Assuming that the dipole shown in Fig. 24.4 has no spheres at the ends, will there be a current in the two short wire segments? If the answer is yes, what do you expect this current distribution to be like?

**Q24.6.** What is the relationship between the phasor current $I$ in the dipole in Fig. 24.4, and the charges $Q$ and $-Q$ on the dipole end spheres?

**Q24.7.** Take a pencil and assume it is a Hertzian dipole. What is its radiation pattern in space like?

**Q24.8.** In the preceding question, define an $E$ plane and an $H$ plane of the radiation pattern.

**Q24.9.** What is an isotropic antenna? Can it be made? If you think it cannot, explain why.

**Q24.10.** Why is the directivity of an isotropic antenna equal to unity, or zero dB?

**Q24.11.** What are the conditions implicit in the definition of the effective antenna area?

**Q24.12.** Can the Friis transmission formula be used for the analysis of a radio communication channel if the transmitting antenna is not matched? Or if the receiving antenna is not matched? How would you modify the formula?

## PROBLEMS

**P24.1.** Prove that the impedance of any antenna above a perfectly conducting ground, with the generator driving the antenna connected between the ground and the antenna terminal, is one half that of the symmetrical antenna obtained with the image of the antenna.

**P24.2.** A thin two-wire transmission line with conductors of radius $a = 1$ mm and distance between them $d = 5$ cm is driven at one end by a generator with a rms value of the emf $\mathcal{E} = 10$ V and frequency $f = 100$ MHz. The line length is $b = 50$ cm, and the other end of the line is open-circuited. Assuming that the line conductors do not radiate, but that the short segment with the generator does, determine approximately the electric field strength at a distance $r = 1$ km from the antenna. (*Hint:* consider the short segment with the generator as a Hertzian dipole.)

**P24.3.** A Hertzian dipole of length $l = 1$ m is fed with a current of rms value $I = 1$ A and of frequency $f = 1$ MHz. Find the rms values of $E_\theta$ and $H_\varphi$ in the equatorial plane (plane $\theta = \pi/2$) of the dipole at a distance of $r = 10$ km.

**P24.4.** Using a system of two half-wave dipoles, construct an antenna system that radiates a circularly polarized wave in one direction. State clearly how you would make the feed.

**P24.5.** A short vertical transmitting antenna of height $h$ has a conducting plate at the top, so that the current along the antenna is practically uniform, of rms value $I$. At the receiving point, at a distance $d$ from the antenna, only the wave reflected by the ionosphere arrives, as shown in Fig. P24.5. The ionosphere can be approximated by a perfectly conducting plane at a height $H$ above the surface of the ground. Assuming that the ground at both the transmitting and receiving point is perfectly conducting, and neglecting the curvature of the earth, determine the rms value of the electric field intensity at the receiving point. The wavelength of the radiated wave is $\lambda$.
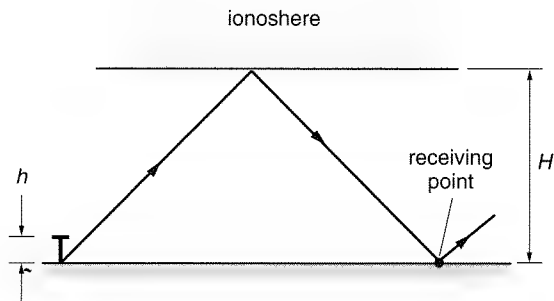


Figure P24.5 Vertical antenna above ground

**P24.6.** In Example 24.7, replace one of the antennas (for example, in a cellular phone, this would be the base-station antenna) by an antenna with a higher directivity and calculate the received power for a directivity of (1) 6 dB, (2) 10 dB, and (3) 20 dB.

**P24.7.** Assume that in a communications link two matched lossless antennas, $A$ and $B$, are $r$ apart in each other's far fields. The antenna directivities and effective areas in the line-of-sight direction are $D_A$, $A_A$, and $D_B$, $A_B$, respectively. First antenna $A$ transmits a power $P_{A1}$, while antenna $B$ receives a power $P_{B1}$. Then antenna $B$ transmits a power $P_{B2}$, while antenna $A$ receives a power $P_{A2}$. Using the reciprocity condition, which says that $P_{A1}/P_{B1} = P_{B2}/P_{A2}$ (think about what this means), show that the ratio of the directivity to the effective area is a constant for any antenna.

**P24.8.** Derive the Friis formula in terms of effective area only. In a microwave relay system for TV each antenna is a reflector with an effective area of $1\,\mathrm{m}^2$, independent of frequency. The antennas are 10 km apart. If the required received power is $P_r = 1$ nW, what is the minimum transmitted power $P_t$ required for transmission at 1 GHz, 3 GHz, and 10 GHz?

**P24.9.** Derive the Friis formula in terms of directivities only.

# 25

# *Some Practical Aspects of Electromagnetic Waves*

## 25.1 Introduction

Applications of electromagnetic waves are numerous, ranging from cooking food to controlling a faraway spacecraft and receiving information from it. Electromagnetic waves cover a very broad frequency (wavelength) spectrum, and it is impossible in this text to even attempt to cover applications in all regions of the spectrum. Therefore, we confine ourselves mainly to waves used in the versatile area of communications, as the readers of this text are likely to spend a part of their professional lives dealing with this subject. The word "communications" refers broadly to sending and receiving a signal that contains some useful information. This signal might be sent along a coaxial cable or through a waveguide, or radiated or received by an antenna, or propagated along an optical fiber, for example. We will describe some issues related to these different ways of communicating, but we will not deal with the information itself. At the end of the chapter, some other common applications of electromagnetic waves, such as cooking, are described briefly.

## 25.2   Power Attenuation of Electromagnetic Waves

When one sends or receives information using electromagnetic (radio) waves, the maximum distance at which this can be done is limited by the amount of power available at the sending end, and the loss of the wave energy by the time it gets to the receiving end, assuming a certain receiver sensitivity. The path loss varies with the medium through which the wave is propagating, as well as the frequency (wavelength) of the wave. So let us calculate the loss per unit distance for a few different cases, and then do a performance comparison. In the following examples, we calculate the loss in a coaxial cable, a rectangular waveguide, an optical fiber, and a line-of-sight radio link (Fig. 25.1), referring to knowledge we gained in previous chapters. In each case, we consider a link where the power at the transmitting (sending) end is $P_T$ and the received power $P(r)$ is a function of the distance $r$ between the transmitter and receiver. So, at a point $r$ away from the transmitter, the received power is $P(r) = P_T f(r) < P_T$. What is this function $f(r)$ in different cases?

**Example 25.1—Attenuation of an electromagnetic wave in a coaxial cable.** In a coaxial cable, if the losses are not large, the power along the line as a function of distance $r$ from the generator can be expressed as

$$P(r) = P_T e^{-2\alpha r} \tag{25.1}$$



(a)          (b)

(c)          (d)
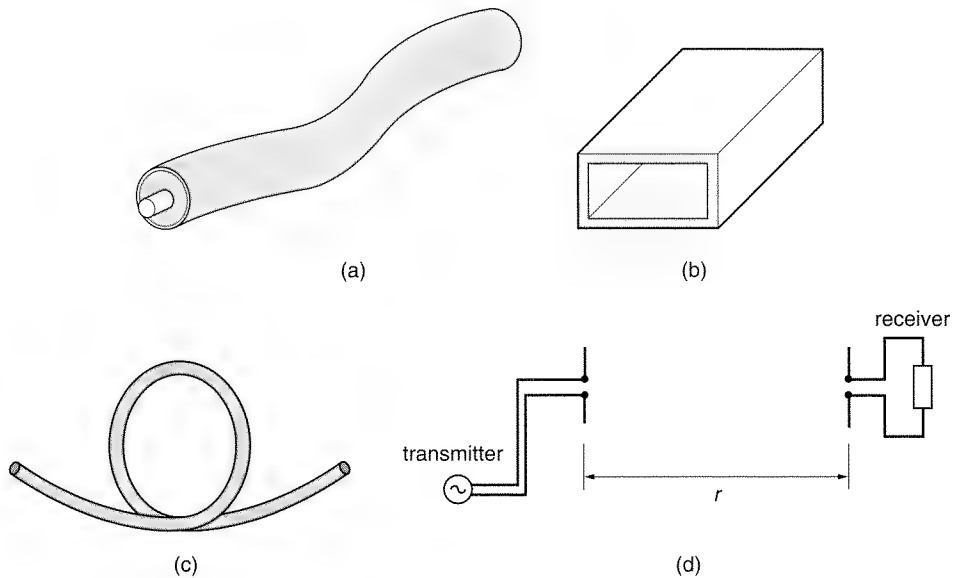
Figure 25.1 (a) A coaxial cable, (b) a rectangular waveguide, (c) an optical fiber, and (d) antennas in a line-of-sight radio link are used in communications as ways of transferring information contained in an electromagnetic wave. The coaxial cable was discussed in detail in Chapter 18, the rectangular waveguide in Chapter 23, and the line-of-sight link in Chapter 24.

(see section 18.4, and note that power is the product of the voltage and current, giving the factor of 2 in the exponent). Therefore,

$$-\frac{\mathrm{d}P(r)}{\mathrm{d}r} = 2\alpha P_T \mathrm{e}^{-2\alpha r} = 2\alpha P(r) = \frac{\mathrm{d}P_{\mathrm{losses}}(r)}{\mathrm{d}r}. \tag{25.2}$$

The attenuation coefficient $\alpha$ comes from losses in the conductor, described by the resistance per unit length, $R'$, and losses in the dielectric, described by the conductance per unit length, $G'$, as described in Fig. 18.10. Usually the conductive losses are dominant, and

$$\frac{\mathrm{d}P_{\mathrm{losses}}(r)}{\mathrm{d}r} = R'|I(r)|^2.$$

The power transmitted to a point $r$ away from the line's beginning can be expressed in terms of the current $I(r)$ and the characteristic impedance (assuming no losses) as $P(r) = Z_0|I(r)|^2$, where $Z_0 \simeq \sqrt{L'/C'}$. So

$$\alpha = \frac{R'}{2Z_0} = \frac{R'}{2}\sqrt{\frac{C'}{L'}}. \tag{25.3}$$

As one example, let us calculate the loss at 1 MHz in a coaxial cable made of copper, filled with a dielectric of permittivity $\epsilon_r = 3$, and of dimensions such that the inner conductor radius $a = 0.45\,\mathrm{mm}$ and inner radius of the outer conductor $b = a\mathrm{e}$. The skin depth is $\delta = 0.067\,\mathrm{mm}$ (Example 20.1), which means that the current is not distributed through the entire cross section. We obtain in this case that $R' \simeq \rho/(2\pi a\delta) = 0.093\,\Omega/\mathrm{m}$, and $Z_0 = 50\,\Omega$. This gives $\alpha = 0.00093\,\mathrm{Np/m} = 0.016\,\mathrm{dB/m}$, so that $f(r) = \mathrm{e}^{-0.0019r}$, where $r$ is in meters. This just corresponds to the loss in the inner conductor. In the outer conductor, the losses are lower (see problem P25.5).

What happens at higher frequencies with the loss in coaxial cables? As another example, let us look at losses at 0.1, 1, and 10 GHz in a high-frequency 50-$\Omega$ RG-58 cable, made of copper with polyethylene dielectric, with $a = 0.45\,\mathrm{mm}$ and $b = 1.47\,\mathrm{mm}$. The loss can usually be found in manufacturer's data sheets, and for this cable at 0.1 GHz, the loss is about 0.2 dB/m, and at 1 and 10 GHz it increases to 0.66 and 2.6 dB/m, respectively. This means that at 10 GHz almost half of the power (which would be 3 dB) is lost after only 1 m of propagation. In this case, the function $f(r) = \mathrm{e}^{-0.6r}$, where $r$ is in meters. Why is the loss this high?

**Example 25.2—History: the transatlantic telegraphy cable. Reduction of losses along lines by increasing inductance per unit length.** When the first transatlantic cable was laid in the Atlantic Ocean, the engineers did not understand that the loss over several thousand kilometers would make the cable impractical (see problem P25.1—calculate the loss for $r = 5000$ km at 10 kHz for practice). The famous British physicist Oliver Heaviside had warned the engineers about losses, but they did not listen. Later, Mihailo Pupin, a Columbia University professor, noticed that in practice, the first term of $\alpha$ in parentheses in Eq. (18.36), repeated here for convenience,

$$\alpha \simeq \frac{1}{2}\left( R'\sqrt{\frac{C'}{L'}} + G'\sqrt{\frac{L'}{C'}} \right). \tag{18.36}$$

is much greater (several orders of magnitude) than the second, due to the relatively large value of $R'$. He next realized that it is possible to reduce this term considerably by increasing $L'$,

without making the second term prohibitively large. He then proposed to reduce $\alpha$ by placing series inductive coils along the cable at regular distances. These are today called Pupin coils, and they enabled transmission of signals along transmission lines of great lengths, including the transatlantic cable.

Let us consider a typical coaxial line and estimate the first and second terms in parentheses of Eq. (18.36). Let the line dielectric have a relative permittivity $\epsilon_r = 3$, permeability $\mu_0$, and conductivity $10^{-12}$ S/m. Let the line conductors be copper, of conductivity $56 \times 10^6$ S/m, and let the ratio of conductor radii (see Table 18.1) be $b/a = e = 2.71828$.

Using the expressions for $C'$, $G'$, $L'$, and $R'$ in Table 18.1, we find that $C' \simeq 167$ pF/m, $G' \simeq 0.3$ pS/m, $L' \simeq 0.2\,\mu$H/m, and $R' \simeq 0.0055\,\Omega$/m. With these values, the first term in the expression for $\alpha$ (including $1/2$) is about $0.8 \times 10^{-4}$ Np/m, and the second term is about $5 \times 10^{-12}$ Np/m. The first term, due to imperfect cable conductors, is indeed much greater than the second, due to imperfect cable dielectric.

By increasing $L'$ artificially, as Pupin did by connecting lumped series coils in the cable, the first term can be substantially reduced, and therefore $\alpha$ made smaller. Note that the attenuation of the wave is proportional to $e^{-\alpha r}$, so a 10-fold increase in inductance per unit length results in about a 24-fold decrease in signal attenuation. This means that if a cable with no Pupin coils has an attainable range of 100 km, with a 10-fold artificial increase of the line series inductance per unit length the range is increased to about 2400 km.

**Example 25.3—Attenuation of electromagnetic waves in a rectangular waveguide for the dominant mode.** Losses in hollow metal waveguides depend on the mode propagating in the waveguide, the type of metal and dielectric used to make the waveguide, the geometry of the waveguide, and frequency. It can be shown (see, e.g., S. Ramo et al., *Fields and waves in communication electronics*, 3d ed., J. Wiley & Sons, 1993, p. 423) that the attenuation coefficient for the dominant $TE_{10}$ mode in a rectangular waveguide of sides $a$ and $b$ is given by

$$\alpha = R_s \sqrt{\frac{\epsilon}{\mu}} \frac{a/b + 2f_c^2/f^2}{a\sqrt{1 - f_c^2/f^2}} \qquad (TE_{10}), \tag{25.4}$$

where $R_s$ is the surface resistance of the metal walls, Eq. (20.10). Consequently, the losses depend on the metal conductivity and frequency. For a typical X-band (8.2 to 12.4 GHz) waveguide ($a = 25.4$ mm, $b = 12.7$ mm), made of brass plated with silver and a rhodium anticorrosion coating, the conductivity is $6.17 \cdot 10^7$ S/m, and the skin depth at 10 GHz is $\delta = 0.64\,\mu$m, yielding $R_s = 2.53\,\Omega$ and $\alpha = 0.0883$ Np/m $= 0.767$ dB/m.

Note that at very high frequencies, when the skin depth is on the order of the conductor surface imperfections (the surface cannot be made absolutely flat), the surface resistance becomes substantially larger than its theoretical value for a perfectly flat surface.

It should be noted that some other kinds of metallic waveguides have field profiles that result in lower losses than in the previous case. An example is a $TM_{11}$ mode in a waveguide with circular cross section, with a typical $\alpha = 0.01$ dB/m. Initially there was a large development effort, mostly in Bell Labs, to use this kind of waveguide for the entire phone network across the United States, but before the network was built, optical fibers were shown to have lower loss and cost, so the waveguide technology was never implemented. In Socorro, New Mexico, however, there is a large radio telescope [Very Large Array (VLA), nicely shown in the 1997 movie *Contact*] in which the signals received from 27 large dish antennas (each 25 m in diameter) formerly were propagated at 44 GHz through a circular waveguide to the operations center that is about 60 km away from the telescope. Recently, the circular waveguide was replaced by optical fibers, but the waveguide is still used as the pipe for the fibers.
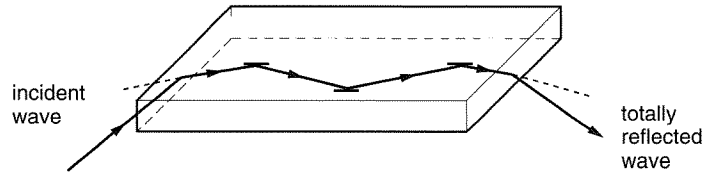
Figure 25.2 Sketch of wave propagation along a dielectric slab

**Example 25.4—Dielectric waveguides and optical fibers.** Optical fibers are used as waveguides for electromagnetic waves in the visible and infrared part of the frequency spectrum, with wavelengths between roughly 300 nm and 10 $\mu$m (optical engineers usually think in terms of wavelength, whereas radio engineers think in terms of frequency). Fibers are so-called *dielectric waveguides*. The simplest dielectric waveguide is a flat dielectric slab. Because the permittivity of the slab is always greater than $\epsilon_0$, a possibility exists that the wave propagating in the slab is totally reflected at the interface.

If we excite in the slab a plane wave incident on one slab face at an angle greater than the critical angle, it will be reflected totally at the same angle toward the other face, then reflected totally from that other face, etc. So the wave will bounce between the two slab faces, and the slab will serve as a guiding medium of the wave, as sketched in Fig. 25.2.

The principle of optical fibers is essentially the same, although the wave types propagating along them are more complicated. Thanks to total reflection, however, these waves also are restricted to the domain of the fiber. The fiber is made of an inhomogeneous dielectric (quartz), and roughly speaking it has a core and an outer cladding layer, sketched in Fig. 25.3. The permittivity of the core is typically a fraction of a percent higher than that of the outer part (it is germanium doped). The cladding has a permittivity of about $\epsilon_r = 2.1$ (an optical index of $n = \sqrt{\epsilon_r} = 1.46$). The typical attenuation of a so-called single-moded fiber at a wavelength of 1.55 $\mu$m is 1 dB/km = 0.001 dB/m (Corning specification sheets, 1998), and for very specialized low-loss fibers it can be as low as 0.1 dB/km.

**Example 25.5—Attenuation of electromagnetic waves in a line-of-sight radio link through a vacuum.** In a line-of-sight radio link (which means that the radio wave travels between two antennas directly, with no reflections), the power loss function $f(r)$ is given by the Friis transmission formula:

$$P(r) = P_T \frac{G_T A_R}{4\pi r^2} = P_T \frac{A_T A_R}{\lambda^2 r^2}, \tag{25.5}$$
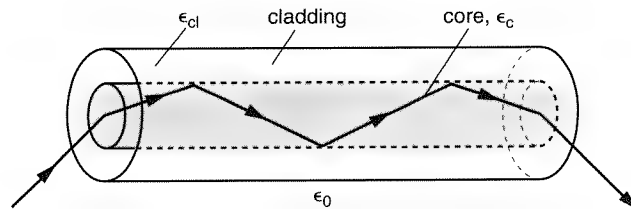


Figure 25.3 Sketch of wave propagation along an optical fiber, based on total internal reflection

where $G_T$ is the gain of the transmitting antenna, and $A_R$ and $A_T$ are effective areas of the receiving and transmitting antennas used in the link. We can measure these effective areas in terms of the operating wavelength $\lambda$ used in the link. If the two antennas are equal, $n\lambda^2$ large, and we assume they are well designed so that the effective areas are roughly equal to their geometric areas, we get

$$\frac{P(r)}{P_T} = f(r) = \frac{n^2\lambda^2}{r^2}. \tag{25.6}$$

This means that the larger the antennas are (measured in wavelengths), the lower the loss of power between the transmitter and receiver. Large antennas, of course, correspond to high directivity (gain). As an example, a standard X-band horn antenna measures about 2 by 2.6 wavelengths at 10 GHz, yielding $f(r) = 0.025/r^2$. If instead of this horn, 3-m round dishes are used, $f(r) = 55,000/r^2$. In the second case, a much larger distance can be spanned with the same receiver sensitivity.

In the preceding examples we have seen that in coaxial cable, waveguides, and optical fiber, the power decays exponentially away from the transmitter, with very different decay constants (on the order of 1 dB/m in a coaxial cable, 0.1 dB/m in a waveguide, and 0.01 dB/m in a fiber). If antennas are used, however, the power drops as $1/r^2$, which is a function that decays more slowly than an exponential for large distances. A plot that shows the attenuation as a function of distance is shown in Fig. 25.4, for the attenuation coefficient values calculated in Examples 25.1 to 25.5. The lower attenuation for long distances is one of the reasons that antennas are used
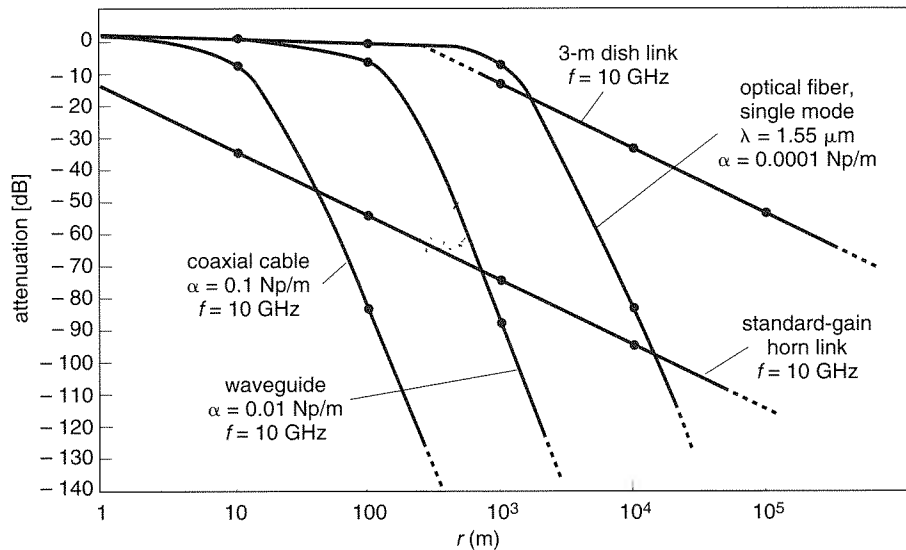


Figure 25.4 The attenuation function $\log f(r)$ for coaxial cable, rectangular waveguide, and a 3-m diameter dish antenna line-of-sight link at 10 GHz. Attenuation in an optical fiber at 1.55-$\mu$m wavelength is also shown. Note the logarithmic scale.

for communications. Another reason is that in many cases, for example in aircraft guidance, satellite communications, portable phones, and pagers, it would not be practical to use cables.

**Example 25.6—Curvature of the earth and effective earth radius in line-of-sight links.** AM broadcasting systems rely on surface wave transmission between two points on the earth's surface. Shortwave radio systems bounce waves off the ionosphere and the earth's surface. The UHF and VHF radio waves used for communications by airplanes, as well as microwaves in radio relay links, propagate along a direct path. As mentioned, this is called *line-of-sight* propagation, illustrated in Fig. 25.5. The figure shows how the range is limited by the curvature of the earth. That is why almost all radio relay stations are put on high peaks, even though the weather conditions at these places often complicate the design (and very few people want to work there). From the figure,

$$(R + h)^2 = R^2 + r^2, \tag{25.7}$$

where $R$ is the radius of the earth, $h$ is the height of the antenna above ground, and $r$ is the range of the communication link. If we solve for $r$,

$$r = \sqrt{h^2 + 2Rh} \simeq \sqrt{2Rh} \tag{25.8}$$

since $R \gg h$.

Equation (25.8) predicts shorter ranges than the ones achievable in reality. The reason is the change in the refractive index of the atmosphere, so that the waves really follow a curved path, not a straight one. This is sketched in Fig. 25.6. The waves bend toward the denser (lower) layers, and this gives longer ranges. This effect varies quantitatively depending on the location on the earth and the hour of the day. A reasonable approximation is to use an effective radius for the earth, $R_{\text{eff}}$, somewhat larger than the real radius. It turns out that if this radius is taken to be about 4/3 of the actual radius, or about 8500 km, the wave refraction is fairly accurately taken into account. If both the receiving and transmitting antennas are above ground, the line-of-sight approximate range formula becomes

$$r = \sqrt{2R_{\text{eff}}h_1} + \sqrt{2R_{\text{eff}}h_2}. \tag{25.9}$$
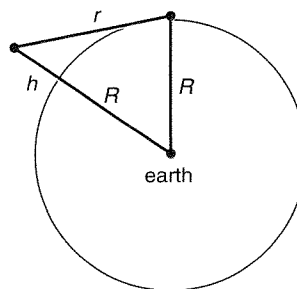


**Figure 25.5** Line-of-sight path limit on the curved earth

path of wave

$h_1$    $h_2$

increasing
index of
refraction $n$
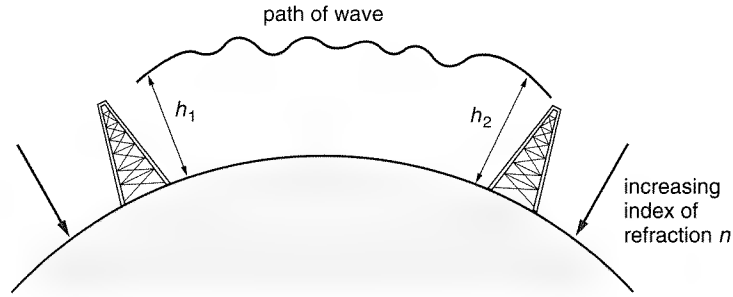
Figure 25.6 The variation of the refractive index of the atmosphere
makes the paths of the waves longer.

When deriving the antenna link path loss, we assumed no extra attenuation due
to the atmosphere. This is a good assumption in clear weather close to the earth's
surface, but only in a certain range of frequencies. Rain and snow degrade the path
loss significantly, but even clear atmosphere has a frequency-dependent attenuation
curve that has strong peaks due to specific properties of oxygen, the hydroxide (OH)
radical, water vapor, and other constituents of the atmosphere. This dependence dic-
tates the frequencies used for specific purposes, as described later in this chapter.
In satellite communications, the waves pass through the ionosphere, a layer of the
atmosphere that has unique properties and that also significantly affects wave prop-
agation. We will be able to analyze the effects of the ionosphere in more detail after
we develop an understanding of plane wave propagation through ionized gases, de-
scribed in the next section.

*Questions and problems:*   Q25.1 to Q25.5, P25.1 to P25.8

## 25.3   Effects of the Ionosphere on Wave Propagation

The upper layer of the atmosphere, between about 50 and 500 km above the earth's
surface, is a highly rarefied ionized gas. This ionized layer of the atmosphere is
known as the *ionosphere*. It has a pronounced influence on the propagation of electro-
magnetic waves in a wide frequency range. It is therefore important to understand
this influence when dealing with any kind of radio communications in which the
waves travel through the ionosphere. The influence of the ionosphere on radio-wave
propagation is of great interest starting from very low frequencies (10 to 100 kHz), to
short waves (up to 30 MHz), but also for higher frequencies than these.

### 25.3.1   PLANE WAVE PROPAGATION THROUGH THE
IONOSPHERE (AN IONIZED GAS)

This section is aimed at presenting the basic theory of propagation of uniform plane
electromagnetic waves in ionized gases. To simplify the analysis, the collisions of
moving charged particles with neutral gas molecules, resulting in wave attenuation,
will be ignored.

Consider a uniform plane electromagnetic wave of angular frequency $\omega$ propagating in an ionized gas. Let there be $N$ ions per unit volume of charge $Q$ and mass $m$. Assume that at a fixed point inside the gas the electric field strength of the wave varies in time as $\mathbf{E}(t) = \mathbf{E}_m \cos \omega t$. The equation of motion of a single ion under the influence of the electric and magnetic field of the wave has the form

$$m\frac{d\mathbf{v}}{dt} = Q\mathbf{E}_m \cos \omega t + Q\mathbf{v} \times (\mu_0 \mathbf{H}_m) \cos \omega t. \tag{25.10}$$

We know that for a uniform plane wave $H_m = \sqrt{\epsilon_0/\mu_0}E_m$, and that $\sqrt{\epsilon_0\mu_0} = 1/c_0$. Therefore, the second term on the right side of this equation is approximately proportional to the first term multiplied by the ratio $v/c_0$. Because the velocities that ions can acquire in a time-harmonic electric field are much smaller than the velocity of light in a vacuum, the second term can be ignored. If we multiply the equation thus obtained by $dt$ and then integrate, we obtain

$$\mathbf{v} = \frac{Q}{\omega m}\mathbf{E}_m \sin \omega t. \tag{25.11}$$

The integration constant, a time-constant velocity, is zero because a time-harmonic electric field cannot produce a steady drift of ions.

Knowing the velocity of ions, we know also the current density that they produce,

$$\mathbf{J} = NQ\mathbf{v} = \frac{NQ^2}{\omega m}\mathbf{E}_m \sin \omega t. \tag{25.12}$$

The second Maxwell's equation now becomes

$$\nabla \times \mathbf{H} = \mathbf{J} + \epsilon_0\frac{\partial \mathbf{E}}{\partial t} = \left(\frac{NQ^2}{\omega m} - \epsilon_0\omega\right)\mathbf{E}_m \sin \omega t, \tag{25.13}$$

or

$$\nabla \times \mathbf{H} = -\omega\left(\epsilon_0 - \frac{NQ^2}{\omega^2 m}\right)\mathbf{E}_m \sin \omega t. \tag{25.14}$$

For $N = 0$ (a vacuum), the second term in the parentheses does not exist. Therefore, the presence of the ions can be represented by an equivalent *reduction* in permittivity. Because this reduction is proportional to $Q^2/m$, we conclude that the sign of the ions is unimportant, and that those ions having the largest ratio $Q^2/m$ have the most pronounced influence. Because this ratio is the largest for free electrons, they are the dominant factor for plane wave propagation in an ionized gas.

Let us define the equivalent (or effective) permittivity of an ionized gas by

$$\epsilon' = \epsilon_0\left(1 - \frac{NQ^2}{\omega^2\epsilon_0 m}\right) = \epsilon_0\left(1 - \frac{\omega_c^2}{\omega^2}\right), \tag{25.15}$$

where

$$\omega_c = \sqrt{\frac{NQ^2}{\epsilon_0 m}} \qquad (25.16)$$

is known as the *critical angular frequency*, and $f_c = \omega_c/2\pi$ as the *critical frequency*, of the ionized gas.

The propagation coefficient can now be written as

$$\beta = \omega\sqrt{\epsilon'\mu_0} = \frac{\omega}{c_0}\sqrt{1 - \frac{\omega_c^2}{\omega^2}}, \qquad (25.17)$$

and the phase velocity of the wave is given by

$$v_{\text{ph}} = \frac{\omega}{\beta} = \frac{c_0}{\sqrt{1 - \omega_c^2/\omega^2}}. \qquad (25.18)$$

If $\omega > \omega_c$, the expression under the square root is positive, so that $v_{\text{ph}} > c_0$. We know that this is only a geometrical velocity, and that the velocity of propagation of a signal, or of energy, is less than $c_0$ (see Example 21.6).

If $\omega < \omega_c$, however, the expression under the square root is negative and $\beta$ becomes imaginary, which means that waves of angular frequencies less than $\omega_c$ cannot propagate in this ionized gas. This is why $\omega_0$ is called "the critical angular frequency," and $f_c$ "the critical frequency."

**Example 25.7—Penetration of plane waves through the ionosphere.** The critical frequency of the ionosphere varies greatly with the distance from the earth's surface, as well as with the hour of the day and month of the year, and with the sun's activity. Roughly, for a plane wave propagating vertically, this frequency ranges from about 3 to 8 MHz. Therefore, no wave of a frequency below about 3 MHz can escape the earth, nor can such a wave reach us from outer space. Therefore, for communications via satellites we must use higher frequencies than these.

It is interesting that at very low frequencies (less than about 100 Hz), the wavelength is much larger than the ionosphere thickness, and such waves can penetrate the ionosphere.

## 25.3.2 REFLECTION AND REFRACTION OF PLANE WAVES IN THE IONOSPHERE

We shall now see what happens if a plane wave is emitted from the earth's surface toward the ionosphere at an arbitrary angle. The ionosphere is an ionized layer of the atmosphere, and we have shown earlier that free electrons have the most pronounced influence on wave propagation. The concentration of electrons changes with the height and depends greatly on the hour of the day, and significantly on the season of the year, the latitude, and the activity of the sun.

During the day, the variation of the concentration of electrons with height above the earth's surface shows certain regularities, resembling four blurred layers. Starting from the surface of the earth, the layers are designated as $D$, $E$, $F_1$, and $F_2$. The corresponding heights are 50 to 70 km (the $D$ layer), 100 to 150 km (the $E$ layer),

about 200 km (the $F_1$ layer), and between 250 and 300 km (the $F_2$ layer). During the night, the $D$ and $E$ layers practically disappear, and the layers $F_1$ and $F_2$ merge into a single layer, $F$, between about 250 and 400 km above earth. The critical frequency of layers $E$, $F_1$, $F_2$, and $F$ are about 3 to 4 MHz, 4 to 5 MHz, 6 to 8 MHz, and 3 to 5 MHz, respectively. The lowest layer, $D$, in which collisions of electrons with neutral atoms and molecules are most pronounced, dominates the attenuation of waves propagating through the ionosphere. This is why attenuation of waves reflected from the ionosphere is the largest during the day, when this layer is present.

Assume that the ionosphere critical frequency versus height $h$ above the earth's surface is as in Fig. 25.7, with a maximum critical frequency $f_{c\,max}$ at a certain height. Assume that an antenna radiates a plane wave of frequency $f$ so that it is incident at an angle $\theta_0$ on the lower boundary of the ionosphere (which we assume to be plane), as in Fig. 25.7. We know that in the case of a homogeneous ionized gas, it can be considered as a medium of equivalent apparent permittivity from Eq. (25.15):

$$\epsilon' = \epsilon_0 \left( 1 - \frac{\omega_c^2}{\omega^2} \right) = \epsilon_0 \left( 1 - \frac{f_c^2}{f^2} \right). \tag{25.19}$$

We can imagine the ionosphere consisting of many thin homogeneous layers of slightly different critical frequencies, as indicated in Fig. 25.7. The incident wave then
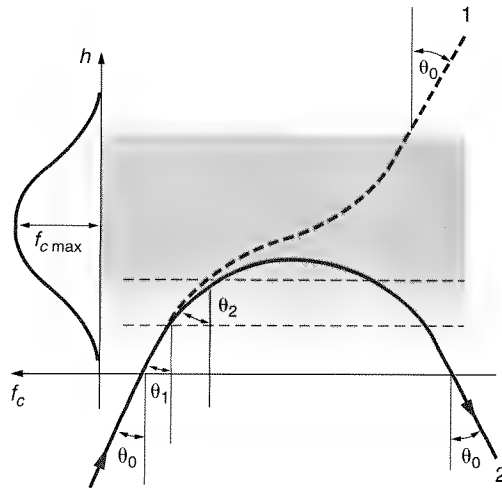


Figure 25.7 A model of the ionosphere critical frequency versus the height, $h$, above earth's surface (diagram on the left), and a sketch of propagation of two waves incident on the ionosphere. The solid line represents a wave of frequency less than the maximum critical frequency. The dashed line represents a wave of frequency greater than the maximum critical frequency.

refracts on the first layer, with no reflected wave because the apparent permittivity of that layer is almost the same as $\epsilon_0$. It is next incident at a new angle, $\theta_1$, on the next layer, and so on. By applying Snell's law, we obtain the following sequence of equations:

$$\frac{\sin \theta_0}{\sin \theta_1} = \sqrt{\frac{\epsilon_1'}{\epsilon_0}} \qquad \frac{\sin \theta_1}{\sin \theta_2} = \sqrt{\frac{\epsilon_2'}{\epsilon_1'}} \qquad \frac{\sin \theta_2}{\sin \theta_3} = \sqrt{\frac{\epsilon_3'}{\epsilon_2'}} \qquad \ldots \tag{25.20}$$

where $\epsilon_1', \epsilon_2', \ldots$, are effective permittivities of successive layers. Let $\theta'$ be the incident angle at any desired height $h$, where the effective permittivity is $\epsilon'$. The angle $\theta'$ can be calculated if we multiply together all the Eqs. (25.20) up to the angle $\theta'$. It is easily seen that the result of this multiplication is

$$\frac{\sin \theta_0}{\sin \theta'} = \sqrt{\frac{\epsilon'}{\epsilon_0}}. \tag{25.21}$$

From this equation and Eq. (25.15) we obtain

$$\sin \theta' \sqrt{1 - \frac{f_c^2}{f^2}} = \sin \theta_0. \tag{25.22}$$

If the frequency $f$ of the wave and the initial incident angle $\theta_0$ are such that $\theta' < \pi/2$ at the height at which the critical frequency $f_c = f_{c\,max}$, the wave will be bent, but *will pass through the ionosphere*, and leave it at exactly the angle $\theta_0$ (case 1 in Fig. 25.7).

If, however, $f$ and $\theta_0$ are such that $\theta' = \pi/2$ before the wave reaches the layer of maximum critical frequency, the wave is reflected from the ionosphere (case 2 in Fig. 25.7), leaving the ionosphere in the downward direction also at an angle $\theta_0$. The wave reaches the height at which the ionization is such that

$$\sqrt{1 - \frac{f_c^2}{f^2}} = \sin \theta_0, \tag{25.23}$$

which, after simple manipulations, becomes

$$f_c = f \cos \theta_0. \tag{25.24}$$

**Example 25.8—Waves incident normally on the ionosphere.** According to Eq. (25.23), for $\theta_0 = 0$ (normal incidence on the ionosphere), $f_c = f$. So, with a wave incident perpendicularly on the ionosphere, all waves of frequencies less than $f_{c\,max}$ will be reflected back. However, all waves of frequencies greater than $f_{c\,max}$ will go through the ionosphere. This conclusion can be used for the experimental determination of $f_{c\,max}$. One would use a variable-frequency transmitter radiating waves vertically, send wave packages of increasing frequency, and listen to the echo. The frequency at which the echo disappears is the maximum critical frequency of the ionosphere at the time and site of the probing.

**Example 25.9—Waves incident obliquely on the ionosphere.** For $\theta_0 > 0$, Eq. (25.24) tells us that all the waves of frequencies less than $f_{c\,max}/\cos \theta_0$ will be reflected back, and those of

frequencies greater than $f_{c\,max}/\cos\theta_0$ will go through the ionosphere. This means that for larger initial angles of incidence, $\theta_0$, higher-frequency waves are reflected from the same ionosphere.

As mentioned, $f_{c\,max}$ is between roughly 3 and 5 MHz, so no wave of a frequency below about 3 MHz can pass through the ionosphere. However, for perpendicular incidence only, waves of frequencies greater than $f_{c\,max}$ pass through it. If the wave is incident at some other angle, frequencies higher than $f_{c\,max}$ will also be reflected. So the ionosphere and the surface of the earth can be used as a kind of duct, or waveguide, along which waves of appropriate frequencies propagate by bouncing back and forth between the ionosphere and earth's surface. This is used in AM radio broadcasting, AND LONG-RANGE SHORT-WAVE RADIO LINKS.

Obviously, for communications with satellites we must utilize frequencies so high that at practically no angle of incidence is the wave reflected from the ionosphere.

*Questions and problems:*   Q25.6 to Q25.8

## 25.4   Choice of Wave Frequencies and Guiding Medium for Different Applications

At lower frequencies, we have seen that the losses in cables are relatively low, and in applications in which the two ends can be physically connected, coaxial cables are often used. An example is cable television, which is distributed over a 75-$\Omega$ coax between about 54 and 88 MHz (channels 2 to 6), and about 100 to 700 MHz (channels 7 to 99). Each analog channel uses 6 MHz of bandwidth. At these relatively low frequencies, waveguides cannot be used in practice—they would be too large (see problem P25.7).

In a cable TV distribution system, as in Fig. 25.8, in each neighborhood a head end station receives the broadcast signal. Antenna links are used for broadcasting in the frequency range of the channels (VHF and UHF). After the signals are received by
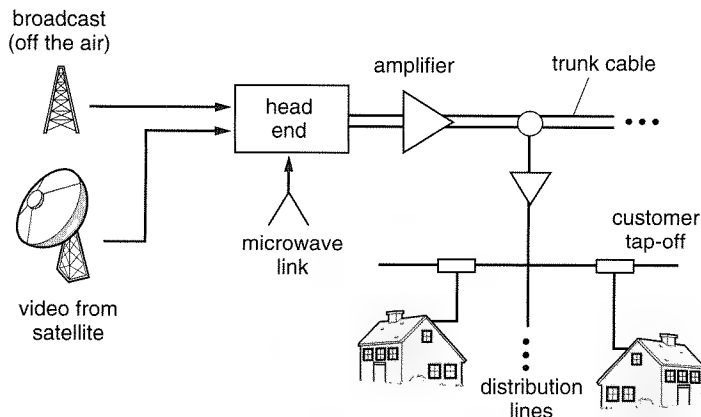


Figure 25.8 Sketch of a cable TV distribution system

one or more antennas, they are first distributed over a trunk cable, and then branched to distribution lines where there is a tap-off for each customer. The cable has loss, so every 40 to 70 m a 20-dB amplifier boosts the signal. As we have seen earlier, the cable will have higher loss at the higher frequencies, so a device called an *equalizer* is used to equalize the power in all the channels.

TV stations can also be received from satellites, in which case the antennas on both ends need to be directional. The channel frequencies are relatively low, so a directional antenna would be quite large (see problem P25.9) and hard to mount on a satellite. In addition, more than one antenna would probably be needed to cover the entire range. As mentioned earlier, to propagate a signal through the ionosphere, a high enough frequency needs to be used so that at practically no angle of incidence is the wave reflected from the ionosphere. The solution to all these problems is to use a higher frequency for the wave transmitted from the satellite to the head station. This is done in such a way that TV channels are translated in frequency, modulating some much higher frequency, which is then radiated from an antenna. On the receiving end, the frequencies are translated back down to the original range. Several properties determine the satellite frequency: size of the antennas, properties of the atmosphere, and available bandwidth.

We discussed antenna size for a given directivity earlier. In satellites, very narrow beam (high-directivity) antennas are used. The satellite is usually several hundred kilometers above the earth's surface, and a narrow beamwidth (corresponding to a small footprint on the surface) translates to electrically large antennas (see problem P25.10). In order for the antenna to fit on a satellite (which is typically a cylinder several meters in diameter and several meters tall), high frequencies (small wavelengths) have to be used.

In addition to frequencies dictated by the ionosphere, the rest of the earth's atmosphere has a pronounced effect on wave propagation. The measured attenuation as a function of frequency at sea level ~~and at a 4-km height~~ is shown in Fig. 25.9. It can be seen that there are some regions with clearly lower attenuation up to about 20 GHz, around 30 to 40 GHz, and around 90 GHz. These are called the *atmospheric windows*. The peaks in the attenuation that define these windows are due to material properties of the different constituents of the atmosphere, as indicated in the figure. Typical frequencies used in satellite communication for TV worldwide are 1.7 to 3 GHz (S band), 3.7 to 4.2 GHz (C band), 10.9 to 11.75 GHz (so-called Ku1 band, although this is really in X band), 11.75 to 12.5 GHz (Ku2 band), 12.5 to 12.75 GHz (Ku3 band), and 18 to 20 GHz (Ka band). Other satellite communications use the regions around 30 GHz and 44 GHz, and some military applications use the 90-GHz region (W band).

In some cases, the high attenuation around 60 GHz is used on purpose. For example, communication between satellites can be done at this frequency with no interference with ground stations. Other examples are wireless local area networks (LANs), which use this frequency because it gives natural cell boundaries due to the high attenuation, as well as some collision avoidance radar systems, which only need to see the nearest obstacles on the road.

The available bandwidth in satellite links determines the amount of information that can be sent. For example, a 6-GHz link with a 5% (300-MHz) bandwidth can
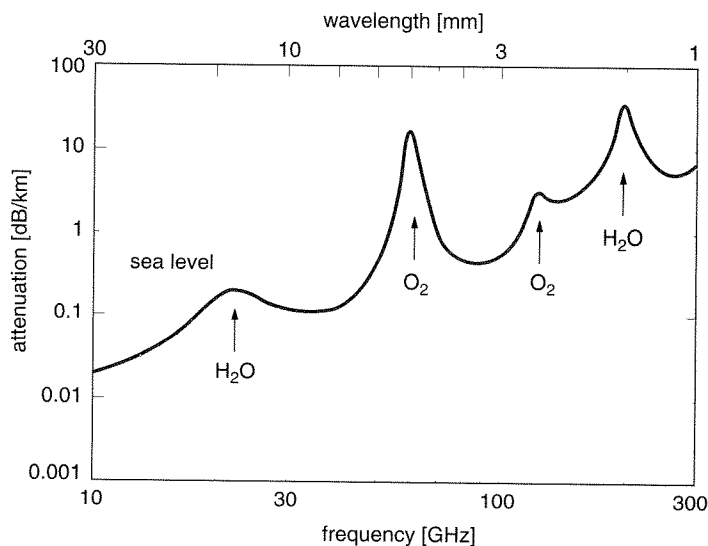
**Figure 25.9** Attenuation of the atmosphere at sea level as a function of frequency. Note the logarithmic scale.

accommodate 50 analog or 25 digital channels. At 30 GHz, a 5% (1500-MHz) bandwidth would accommodate 250 analog or 125 digital channels. For comparison, in an optical fiber at 1.55-$\mu$m wavelength (a frequency of about 200 THz), a 5% bandwidth is very large—over 10 GHz. This large available bandwidth is the major reason for using optical fibers. A standard in 1998 for channels over fibers is a bandwidth of 2.5 Gbit/sec with up to 40 channels (a total bandwidth of over 100 GHz). Another advantage is that one does not have to worry about the atmosphere. The specific wavelength is chosen because very low-loss and low-dispersion fibers can be made at this wavelength.

A number of commercial applications exist for cellular telephony, mostly around 900 MHz and 2 GHz. There are several reasons for choosing these frequencies. The lower part of the spectrum is very noisy due to man-made noise. As an example, the level of man-made noise at 100 MHz is 30 dB lower than at 10 MHz and continues to decrease with frequency. On the upper end, the atmospheric loss due to rain and snowfall increases dramatically above about 3 GHz (as an example, the attenuation in heavy rain is about 0.02 dB/m at 3 GHz, and 2 dB/m at 20 GHz, and a typical cell size is 5 to 10 km). An interesting part of man-made noise is so-called emissions noise from cars: the spark that ignites the combustible mixture of gasoline vapor and air is very nonlinear and has very high harmonics with power levels that are quite high around 2 MHz, but about 40 dB lower at 100 MHz.

In a cellular system, the propagation path is often not direct because the waves bounce off buildings, the ground, and other objects. The situation is also complicated by the fact that users are mobile. Often several waves reach the receiver at the same time, and they might interfere so that they add up or possibly subtract, depending on their relative phases. We will see in the following simple example that with only one
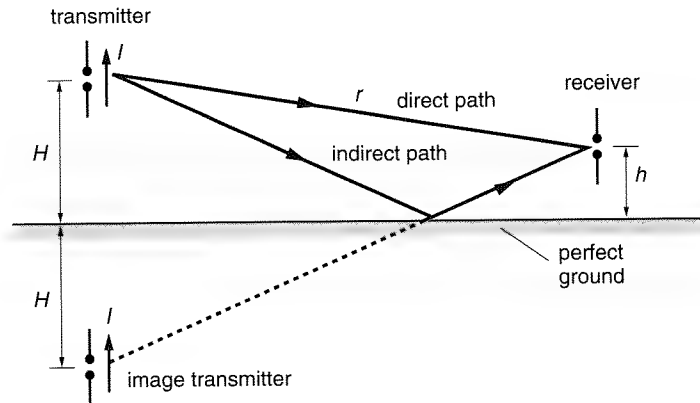
Figure 25.10 A line-of-sight link above a conductive ground: a wave that bounces off the ground interferes with the direct wave, and it can make the received signal smaller or larger. This is the simplest case of so-called multipath fading, which is common in mobile communications.

reflector (the ground), one finds periodic positions where a receiver will detect peaks and nulls. This is called *multipath fading* and exists in all realistic wireless systems, being especially pronounced in mobile systems.

Example 25.10—Line-of-sight link with real ground: a simple multipath fading model. Consider the line-of-sight link shown in Fig. 25.10. The transmitting and receiving antennas are separated by a distance $r$ and are at heights $H$ and $h$ above ground, which is assumed to be a perfect conductor. The receiving antenna receives not only the direct signal but also signals radiated by the transmitting antenna toward the ground that reflect toward the receiver. The effect of the ground can be taken into account by an image of the transmitting antenna, as shown in the figure. At a mobile receiver, the phases of the direct and reflected signals can differ by an even number of half wavelengths, which amounts to the waves adding up or canceling out periodically as the receiver moves away or toward the transmitter. This multipath fading becomes significantly more complicated when other reflective bodies, such as buildings and vehicles, are part of the propagation path.

We use free-space wave propagation every day in a number of places without really thinking about it, for example garage door openers (which typically work at 140 or 450 MHz), or remote controls for home entertainment equipment (which operate in the infrared region with wavelengths between 780 and 860 nm). But sometimes we would like to send information using waves that propagate through a medium other than air, and the issues involved can be very different, which Example 25.11 illustrates.

Example 25.11—Radio communication with submarines. For radio communications in normal circumstances we use frequencies greater than 100 kHz. Assume that we would like to establish a radio link with a submerged submarine. Table 20.1 tells us that this is not possible. Therefore, submarines use much lower frequencies (on the order of 10 kHz) for radio communications. Even this is not sufficiently low, for a submarine must be quite close to the surface in

order to make use of even such low frequencies. The low frequency implies a small bandwidth for communication, which means that very few words per minute can be transmitted.

*Questions and problems:*   Q25.9, P25.9, and P25.10

## 25.5   Radar

Radars are essentially a type of wireless communication link, where the transmitter and receiver are located at the same place, as in Fig. 25.11. The transmitter sends a wave, which eventually reflects off some object (called a *target*, or *scatterer*) partly in the direction from which the wave came. This ("scattered") wave is received at the position of the transmitter, and from it some conclusions can be made about the object that caused the reflected wave.

Radar was invented for military purposes by the British in the Second World War and contributed greatly to the victory of the Allied forces. The word *radar* is an acronym for RAdio Detection And Ranging. Today, there are a number of commercial radar applications, such as weather radar for meteorology, mapping radar, police radar, anticollision radar for cars, and space-imaging radar.

The basic principle of a radar is as follows. The radar transmitter sends a wave with power $P_T$ toward a target. At the target, the power density is $P_T D/(4\pi r^2)$, where $r$ is the distance to the target, and $D$ is the radar antenna directivity. The target scatters the wave proportionally to a quantity called the *radar scattering cross section*, usually denoted by $\sigma(\theta, \phi)$, which is essentially the effective area of the target acting as a receiving antenna. When it reflects the wave, the target acts as a transmitting antenna with a directivity of $4\pi\sigma/(\lambda^2)$. Now the Friis formula can be applied one more time to obtain the power received by the radar receiver:

$$P_{\rm rec} = P_T \frac{D^2(\theta, \phi)\sigma^2(\theta, \phi)}{16\pi^2 r^4}. \tag{25.25}$$

This was derived assuming the radar uses the same antenna for transmission and reception, which is commonly but not always the case (see problem P25.11). The
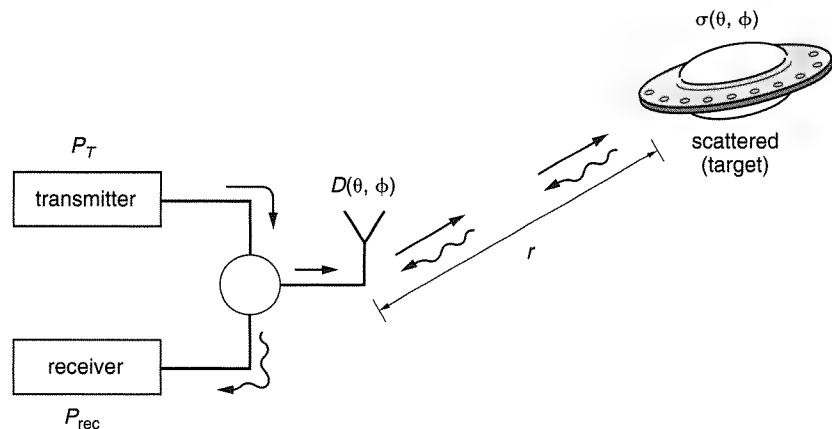


Figure 25.11 A simple schematic of radar operation

received signal in radar is very small, as it falls off as the fourth power of the distance from the target. However, much can be deduced from these signals when properly amplified. Two cases are described in Example 25.12.

**Example 25.12—FM ranging radar and Doppler radar.** In a type of ranging radar, the frequency of transmission is changed linearly from $f_1$ to $f_2$, as in Fig. 25.12. The transmitted signal in this case is said to be *frequency modulated* (FM). If $f_1$ is transmitted, by the time the wave at this frequency reflects back and reaches the receiver, the transmitter is transmitting a different frequency $f < f_2$. In the radar circuit, a so-called beat signal is made, corresponding to the difference $f - f_1$. This difference is obviously dependent on how far the target is, or how long it takes a wave traveling at the speed of light to get there and back. This time is exactly the same time it takes the transmitter to get from $f_1$ to $f$, and is known. So the distance from the target (the range) is

$$r = \frac{cT}{2} \frac{f - f_1}{f_2 - f_1}.$$
(25.26)



(a)



(b)
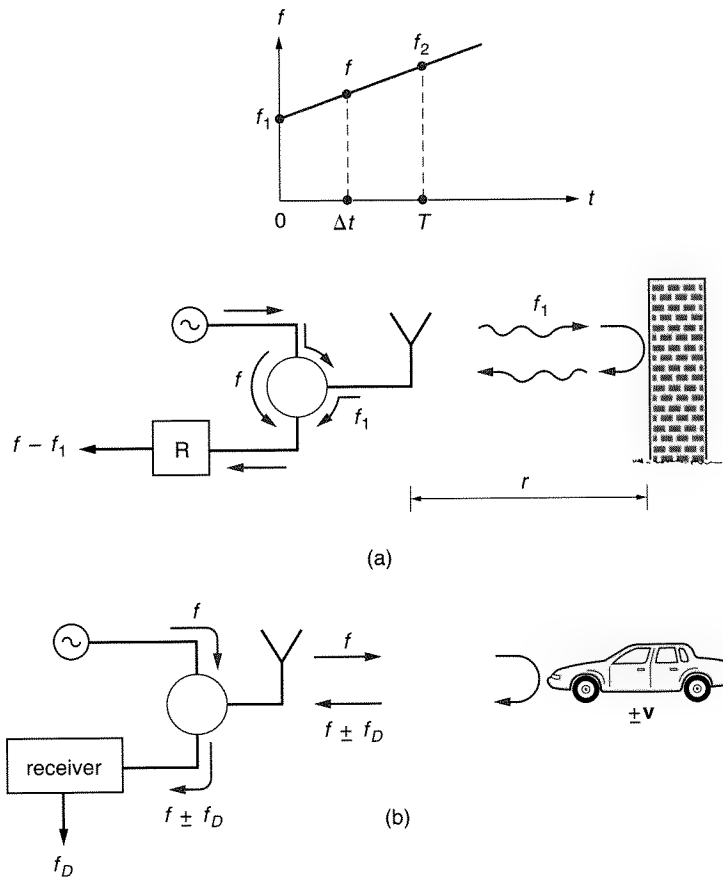
Figure 25.12 (a) Sketch of an FM ranging radar and (b) Doppler radar for measuring speed

This is true for a stationary target. However, if the target is moving, it shifts the received frequency due to the Doppler effect. Using this Doppler shift, the speed of the target can be determined with a radar, as in Fig. 25.12b. In this case, a wave at frequency $f$ is transmitted, and the wave reflected off the target is $f \pm f_D$, where $f_D$ is the Doppler shift, and the sign in front of it depends on whether the target is moving toward or away from the radar. In the receiver circuit, the difference between the two frequencies is measured, and using that, the speed of the target is calculated. Police radars that monitor speed on the roads operate in this way, usually using frequencies around either 10 GHz or 30 GHz.

*Questions and problems:*   Q25.10 to Q25.12, P25.11 to P25.13

## 25.6   Some Electromagnetic Effects in Digital Circuits

We have so far not mentioned wave effects in computers or other digital systems. As the clocks that determine processing speed in computers become faster, electromagnetic effects such as radiation and coupling become more pronounced.

Digital circuits often have printed microstrip transmission lines connecting pins of two chips, possibly through some extra interconnects. The designer wants to make sure that a "one" is indeed a "one" when it reaches the second chip, and the same for a "zero" level. As rise times increase, depending on the logic family, transmission-line effects like overshoot, undershoot, ringing, reflections, and cross talk, can all become critical to maintaining noise margins. For example, in transistor-transistor logic (TTL), the values for a "one" are between 2.7 and 2 V, for a "zero" they are between 0.5 and 0.8 V, and the noise margin is 0.7 V (with a 10 to 90% rise time of 4 to 10 ns). In very fast gallium arsenide (GaAs) digital circuits (with a rise time of 0.2 to 0.4 ns), the values for a "one" are between $-0.2$ and $-0.9$ V, for a "zero" they are between $-1.6$ and $-1.9$ V, and the noise margin is 0.7 V. From these numbers it is seen that digital circuit margins are quite forgiving. For example, in the latter case, a 0.7-V undershoot on a 1.7-V signal is a 45% undershoot, which is considerable. However, it is easy to have a 45% variation in a signal if there is a discontinuity (impedance mismatch) in
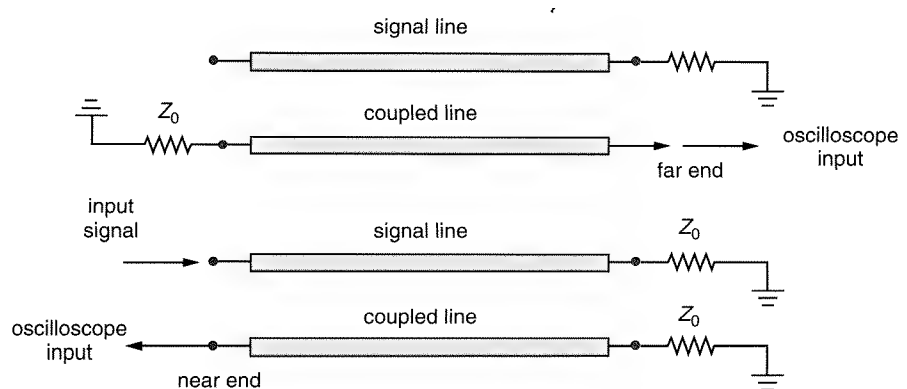


Figure 25.13 Near- and far-end cross-talk measurements in the case of two adjacent printed-circuit board traces

a transmission-line trace on a printed-circuit board. A special type of mismatch is coupling between adjacent traces, which are in effect parasitic inductances or capacitances. This is illustrated in Fig. 25.13 with two lines. If the line labeled "signal line" is excited by a step function, some of the voltage will get coupled to the next closest line even if the line is open-circuited. The coupled signal will appear at both ends of the line, and this is called near- and far-end cross talk. For example, the cross talk could be as high as 25% for two parallel traces, and resistors and trace bends can cause up to 15% and 5% reflections, respectively. All of these could be easily measured with time domain reflectometry (TDR) during the design of the digital backplane (printed-circuit board containing all the traces for chip interconnects).

## 25.7 Cooking with Electromagnetic Waves: Conventional Ovens and Microwave Ovens

In conventional ovens, heating of food is done principally by infrared radiation from the heaters. The infrared electromagnetic region of the spectrum is roughly between 900 nm and 10 $\mu$m. Another way to cook food is with a lower frequency in the microwave region with a wavelength on the order of centimeters. The two cooking mechanisms are quite different because of different skin depths of most foods in the two frequency ranges.

The frequency of infrared radiation is much higher than the highest frequency in Table 20.1. Most often, the food baked in the oven has a conductivity less than that of seawater, but for infrared frequencies the skin depth remains extremely small. We see therefore that a regular oven heats up a very thin surface layer, and then this heat is transferred by *thermal conduction* to the deeper layers. The thermal conductivity of most foods is not high. Therefore, cooking in regular ovens takes a lot of time, in particular if large chunks of food are being cooked. To expedite the process, we use higher temperatures, which result in some drying and browning of the food (at least the parts close to the surface).

In microwave ovens, the standard frequency is 2.45 GHz. Table 20.1 tells us that at that frequency, the skin depth for food (with conductivity usually less, and often significantly less, than that of seawater) is still relatively large (at least 1 cm, and often much larger). Consequently, the microwave oven *instantly* starts to heat most of the volume of the objects in it. Therefore, preparing food is much faster in a microwave oven than in a regular oven, but the food may not brown on the outside. If the cooking time is excessive, much of the water from the food can evaporate, and it can be first dried, then burned (as most of us have probably noticed).

## QUESTIONS

Q25.1. Explain what the physical origin of loss in coaxial waveguides is.

Q25.2. Explain what the physical origin of loss in metallic waveguides is, and why the loss can be smaller than in coaxial cables.

**Q25.3.** Explain what the physical origin of loss in optical fiber is, and why the loss can be smaller than in metallic structures.

**Q25.4.** Explain what the physical origin of loss in a line-of-sight antenna link is.

**Q25.5.** What is the range in a line-of-sight link limited by?

**Q25.6.** Explain in your own words why there is attenuation in an ionized medium with neutral gas molecules.

**Q25.7.** A wave of frequency higher than the highest critical frequency for the ionosphere needs to be used for communication between two points of the earth. Is this possible? Explain.

**Q25.8.** A wave of extremely low frequency (e.g., below 100 Hz) coming from outer space penetrates through the ionosphere and reaches the earth's surface. Explain.

**Q25.9.** Imagine a line-of-sight link in a hallway with conducting walls on top and bottom, and absorbing walls on the sides. How many waves can contribute to the received signal? How would you construct antenna images that approximate the influence of the walls?

**Q25.10.** Derive the radar equation (25.25).

**Q25.11.** Consider a Doppler radar at 10 GHz. The received signal from one car is in the audio range and can be between 300 Hz and 4 kHz. What is the range of speeds this radar can detect?

**Q25.12.** Consider an FM ranging radar in which the frequency varies linearly from $f_1 = 10$ GHz to $f_2$ in $T = 10\,\mu$s. How would you choose $f_2$ in order to be able to detect targets 1 km away, if the radar bandwidth is 500 MHz?

## PROBLEMS

**P25.1.** Calculate how much power is received in England if 1 MW is sent from Boston along a transatlantic 50-$\Omega$ cable at 10 kHz. You can assume that the main loss in the cable is due to conductor loss, and that $R' = 0.005\,\Omega/$m.

**P25.2.** What value of Pupin coils would you choose and how would you place them to reduce the loss in the cable in Example 25.1?

**P25.3.** Calculate the skin depth and attenuation coefficient of a rectangular waveguide with dimensions $a = 23$ mm and $b = 10$ mm, at 10 GHz, if the waveguide is made of (1) copper, (2) aluminum, (3) silver, or (4) gold. What do you think are the engineering problems associated with each metal? Can you think of any combined solution?

**P25.4.** Calculate the skin depth of gold in the optical domain, at wavelengths of 500 nm, 830 nm, 1.33 $\mu$m, and 1.55 $\mu$m. How thin would one need to make a sheet of gold to see through?

**P25.5.** Compare the loss in the inner conductor and outer conductor of a coaxial cable at 1 MHz. Assume the conductors are made of copper, that the cable is filled with a dielectric of permittivity $\epsilon_r = 3$, and that the dimensions are such that the inner conductor radius $a = 0.45$ mm and inner radius of the outer conductor $b = ae$.

**P25.6.** Plot the power attenuation in dB versus distance from 1 m to 1000 km on a logarithmic scale for: coaxial cable at 10 GHz with $\alpha = 0.5$ dB/m, waveguide with $\alpha = 0.1$ dB/m, 1.55-$\mu$m single-mode optical fiber with $\alpha = 0.1$ dB/km, and a free space

link at 10 GHz with a horn antenna with 20-dB directivity and a 1-m diameter dish antenna.

**P25.7.** Calculate the dimensions for a rectangular waveguide with a dominant $TE_{10}$ mode at cable TV frequencies between 100 and 600 MHz.

**P25.8.** A UHF radio system for communication between airplanes uses antennas with a directivity of 2. What is the maximum line-of-sight range between two airplanes at an altitude of 10 km? If the required received power is 10 pW, what is the minimum transmitted power $P_t$ required for successful transmission at 100 MHz, 300 MHz, and 1 GHz?

**P25.9.** Calculate the effective area of a dish antenna for TV that requires a 1-degree beam-width in both $\theta$ and $\phi$ planes, assuming one of the standard cable frequencies (e.g., 225 MHz). Is this a practical antenna? (Note: you can use an approximate formula for the maximal directivity given the beamwidths, $\alpha_1$ and $\alpha_2$, in the two planes, $D \simeq 32,000/(\alpha_1\alpha_2)$, where the beamwidths are given in degrees.)

**P25.10.** If a satellite is 1000 km above the earth's surface, and has a 0.1-degree beamwidth in both planes, calculate the corresponding directivity using the approximate formula in the previous problem. Find the size of the footprint on the earth's surface, and the effective area of the antenna at a satellite frequency of 4 GHz.

**P25.11.** Derive the radar equation (25.25) for a radar that uses two antennas, one for transmitting and another for receiving.

**P25.12.** Assuming a 10-GHz police radar uses an antenna with a directivity of 20 dB (standard horn), and your car has a scattering cross section of $100\lambda^2$, plot the received power as a function of target distance, for a transmitted power of 1 W. If the receiver sensitivity is 10 nW, how close to the radar would you need to slow down to avoid getting a speeding ticket?

**P25.13.** How large is the dynamic range of the radar from problem P25.12? (The dynamic range is the ratio of the largest to smallest signal power detected, expressed in decibels.)

# Appendix 1

# A Brief Survey of Vectors and Vector Calculus

## A1.1 Introduction

If only one number on an appropriate scale is needed to describe a certain quantity, the quantity is known as a *scalar*. For example, the mass of a body is a scalar described by a single number (e.g., $m = 1.6\,\text{kg}$).

Some quantities need a single number to describe them at any one point in a region of space; thus they need an infinite set of numbers to describe them at all points of the region with respect to a scale. For example, air pressure or temperature inside a room can be specified at all points of the room. These quantities are scalars, but to distinguish them from ordinary scalars that need just a single number for their measure, we say that they represent a *scalar field*. A scalar field is described mathematically by a function of three spatial coordinates (and possibly of time, as for a changing temperature in a room). For example, the function $T(x, y, z) = T_0(1 + x + y^2 + z^3)$ defines a specific temperature field.

Many quantities need more than one number to describe them. An important class of quantities is that which needs *three* numbers to describe them completely. These are known as *vectors*. For example, a straight path traversed by a person from one point to another is a vector. It is described by its length (the *magnitude* of the

499

vector), the line along which the motion takes place, and the direction along that line. Other vectors include velocity, acceleration, force, the electric field vector, and the magnetic field vector. In handwriting, vector quantities are denoted by an arrow above the symbol, for example, $\vec{s}$. In printed text, the boldface type is used much more frequently, for example, **s**. For example, we could write

$$\mathbf{s} = \text{New York} \ \rightarrow \ \text{Chicago,}$$

which is unconventional but correct. The magnitude of a vector is denoted by the same symbol, but without the arrow, or not boldface. Sometimes the absolute value sign is used instead, so that $s$, $|\vec{s}|$, and $|\mathbf{s}|$ mean the same: the magnitude of the vector **s**.

Like scalars, many vectors need to be defined by an infinite set of three numbers, e.g., the velocity of all air particles in a region during windy weather. Such vector quantities represent a *vector field*. Vector fields are described mathematically by a *vector function* of coordinates (and possibly of time). For example, in rectangular coordinates, a vector field could take the form $\mathbf{v}(x, y, z, t)$.

*Questions and problems:*   QA1.1 and QA1.2

## A1.2   Algebraic Operations with Vectors

Five basic algebraic operations are used for vector quantities: vector addition, vector subtraction, multiplication of a vector with a scalar (resulting in a vector), multiplication of two vectors resulting in a scalar (known as the scalar, or dot, product), and multiplication of two vectors resulting in a new vector (known as the vector, or cross, product).

### A1.2.1   ADDITION AND SUBTRACTION OF VECTORS

Let us adopt the straight-line motion from one point to another as the prototype of a vector quantity. The vector **A** that describes this displacement is represented as a straight-line segment from its starting point (point 1) to its end point (point 2) (Fig. A1.1a). The arrowhead is added at the vector endpoint to symbolize its direction. Vector **B**, representing another displacement with a starting point at 2 and the end point at 3, is also shown. The total displacement is as if we moved from point 1 to point 3. Vector **C** in Fig. A1.1a, from point 1 to point 3, is therefore defined as the sum of the vectors **A** and **B**, which is written as

$$\mathbf{C} = \mathbf{A} + \mathbf{B}. \tag{A1.1}$$

The vector in Fig. A1.1a directed from point 2 to point 1 is defined as the *negative* of the vector **A**, that is, $-\mathbf{A}$, and that from point 3 to point 2 as the negative of the vector **B**, that is, $-\mathbf{B}$. From the definition of vector addition in Eq. (A1.1) it follows that the vector **D** in Fig. A1.1a represents the difference between vectors **A** and **B**:

$$\mathbf{D} = \mathbf{A} + (-\mathbf{B}) = \mathbf{A} - \mathbf{B}. \tag{A1.2}$$
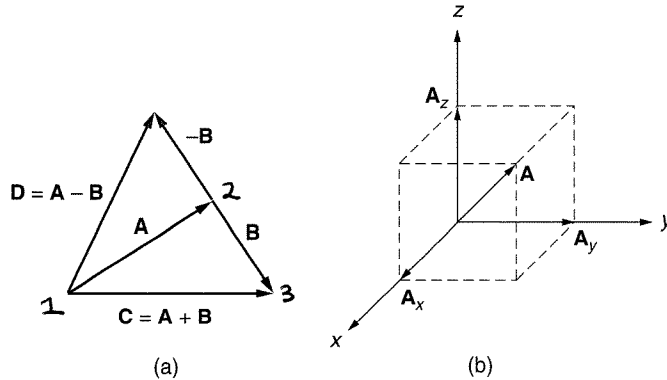
Figure A1.1 (a) Graphical representation of vectors and of vector addition and subtraction. (b) Vector **A** represented as a sum of its three components in the rectangular coordinate system.

Additions and subtractions of more than two vectors are performed two vectors at a time, thus reducing the problem to that of two vectors. A vector in space can always be represented uniquely as a vector sum of three vectors. If these are parallel to the coordinate axes, they are known as the *three vector components* of the vector considered. As an example, vector **A** in Fig. A1.1b is represented by its three vector components in a rectangular coordinate system.

**Example A1.1—Sum of vectors forming a polygon.** Consider first three vectors forming a triangle in the head-to-tails arrangement. To obtain the sum of the three vectors, we first make the vector sum **A** + **B** and find that it is equal to −**C**. Thus the sum of the three vectors is zero.

It is trivial to extend this reasoning to any number of vectors in the head-to-tail arrangement that form a closed polygon. Consider now the limiting case of a polygon with an infinite number of sides formed by differential vector lengths d**l** in the head-to-tail arrangement. The sum of these differential length vectors is a line integral around the contour C they form, and it is equal to zero:

$$\oint_C d\mathbf{l} = 0.$$

Note that the integral of the *magnitudes* of the differential vectors is not zero—it equals the contour length.

## A1.2.2   MULTIPLICATION OF A VECTOR WITH A SCALAR

Multiplication of a vector **A** with a scalar $s$ results in a vector of the same direction as **A**, but of magnitude $sA$. To divide a vector with a scalar amounts to multiplying the vector by the reciprocal of the scalar. For example, $\mathbf{A}/s = (1/s)\mathbf{A}$.

### A1.2.3    UNIT VECTORS

Of particular importance is the concept of the *unit vector*. This is a vector of magnitude one and a pure number (with no physical dimensions). Because we know its magnitude, to specify it, we need only two pieces of information: its direction and sense along that direction. Thus a unit vector defines an *oriented direction in space*. Several notations are used for unit vectors. For example, the unit vector in the direction $x$ may be denoted $\widehat{x}$, $\mathbf{a}_x$, $\mathbf{i}_x$, or $\mathbf{u}_x$. In this text, the symbol $\mathbf{u}$ ("unit vector") with appropriate subscripts is adopted, and sometimes the symbol $\mathbf{r}_0$ for a unit vector in the direction of a radius vector (a vector directed from a fixed point to any direction).

The unit-vector concept is extremely important and can be used on many occasions. Note that *any* vector can be represented as a product of its magnitude and a unit vector, and that the unit vector in the direction of a vector $\mathbf{A}$ is obtained by dividing $\mathbf{A}$ by its magnitude:

$$\mathbf{A} = A\mathbf{u_A}, \qquad \mathbf{u}_A = \frac{\mathbf{A}}{A}. \qquad\qquad (A1.3)$$

*(Relationship between a vector and the unit vector in its direction)*

### A1.2.4    THE DOT PRODUCT OF TWO VECTORS

The *scalar product*, or *dot product*, of two vectors, such as $\mathbf{A}$ and $\mathbf{B}$, denoted as $\mathbf{A} \cdot \mathbf{B}$, is defined as follows. Let the angle between the two vectors be $\alpha$, as in Fig. A1.2. Then

$$\mathbf{A} \cdot \mathbf{B} = AB \cos \alpha. \qquad\qquad (A1.4)$$

*[Definition of dot (scalar) product]*
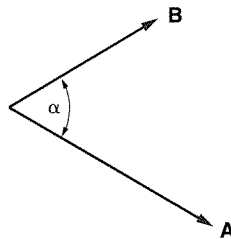
Figure A1.2 The dot (scalar) product of vectors **A** and **B** is defined as $AB \cos \alpha$.
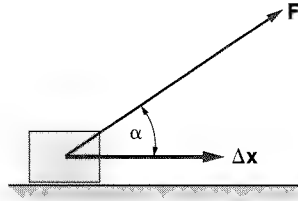
Figure A1.3 The work done by the force **F** in moving the body by $\Delta$**x** is **F** $\cdot$ $\Delta$**x**.

Note that the dot product is a *commutative operation* (i.e., the two vectors in the product can exchange places without affecting the result):

$$\mathbf{B} \cdot \mathbf{A} = BA \cos \alpha = \mathbf{A} \cdot \mathbf{B} = AB \cos \alpha. \tag{A1.5}$$

Finally, the dot product of a vector with itself results in the square of the vector magnitude:

$$\mathbf{A} \cdot \mathbf{A} = A^2. \tag{A1.6}$$

**Example A1.2—Work of a force expressed as the dot product.** Consider a force **F** acting on a body (Fig. A1.3). Assume that the body is free to move only along the $x$ axis. If the body is moved by a small distance $\Delta x$, this displacement can be due only to the projection $F_x$ of the force **F** on the $x$ axis. Therefore the work done by the force **F** is $\Delta A_{\text{of force } \mathbf{F}} = F_x \, \Delta x \cos \alpha$.

According to the definition of the dot product, this can be written in a compact form, $\Delta A_{\text{of force } \mathbf{F}} = \mathbf{F} \cdot \Delta \mathbf{x}$.

**Example A1.3—Dot product of a vector with a unit vector.** Let **u** be the unit vector in an arbitrary direction. The dot product of a vector **A** with the unit vector **u** is given by

$$\mathbf{A} \cdot \mathbf{u} = \mathbf{u} \cdot \mathbf{A} = A \cos(\mathbf{A}, \mathbf{u}) = A_u,$$

where $\cos(\mathbf{A}, \mathbf{u})$ is the cosine of the angle between the two vectors, and $A_u$ is the projection of the vector **A** onto the direction of the unit vector **u**. This is a simple but important result.

## A1.2.5   THE CROSS PRODUCT OF TWO VECTORS

The *cross product* (or vector product) of two vectors **A** and **B** is defined to be a new vector, normal to the plane of vectors **A** and **B**, of magnitude $AB \sin \alpha$ (Fig. A1.4). The cross product is written in the form $\mathbf{A} \times \mathbf{B}$. Thus the magnitude of the cross product in Fig. A1.4 is

$$|\mathbf{A} \times \mathbf{B}| = AB \sin \alpha, \tag{A1.7}$$

*(Magnitude of cross product of two vectors)*
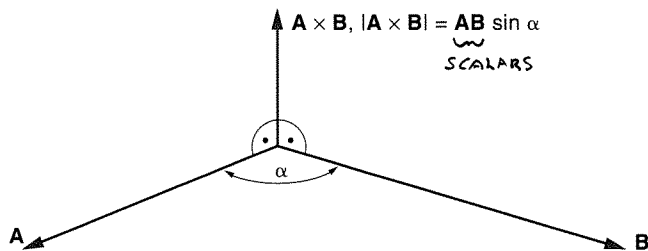
Figure A1.4 Cross product of vectors **A** and **B**, that is, **A** × **B**. Note that **B** × **A** = −**A** × **B**.

and its direction is as indicated in the figure. It is obtained by the *right-hand rule*: if the first vector (**A** in this case) is rotated to coincide with the second vector (**B** in this case) by the smaller of the two angles made by the two vectors, and we imagine rotating our right hand with it, the direction of the cross product is that of the direction of the thumb (or, equivalently, the progression of a right-hand screw). From Fig. A1.4, we have

$$\mathbf{B} \times \mathbf{A} = -\mathbf{A} \times \mathbf{B}, \tag{A1.8}$$

i.e., *the cross product is not a commutative operation.*

**Example A1.4—Moment of force expressed as the cross product.** Consider a force **F** acting on a lever of vector length **r** that cannot move but can rotate only about the $z$ axis, to which it is normal (Fig. A1.5). In rotating the lever, only the component normal to the lever and to the $z$ axis is of interest. The strength with which the lever is forced to rotate is proportional to the lever length and to that force component, and is known as the *moment of force* **F** *with respect to the z axis*. Since this component of the force equals $|\mathbf{F}| \sin \alpha$, where $\alpha$ is the angle between vectors **F** and **r**,

Moment of force **F** with respect to $z$ axis $= |\mathbf{r}|\,|\mathbf{F}| \sin \alpha = |\mathbf{r} \times \mathbf{F}|$

The moment of force is defined to be a vector,

$$\mathbf{M} = \mathbf{r} \times \mathbf{F}, \tag{A1.9}$$

so that by definition, in Fig. A1.5 it is in the direction of the $z$ axis.
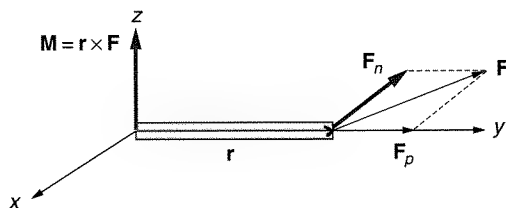


Figure A1.5 The moment of force **F** with respect to the $z$ axis is defined as **r** × **F**. It is assumed that **F** is in the plane normal to $z$.
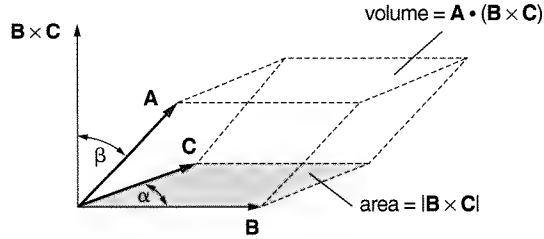
Figure A1.6 Geometrical illustration of the triple scalar product

## A1.2.6   THE TRIPLE SCALAR PRODUCT

We will also need to find products of three vectors. There are two such product types. One is of the form $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$, and it is called a *triple scalar product*. The triple scalar product has the property that

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}), \tag{A1.10}$$

which can be proved as follows. According to the definition of the cross product, the magnitude of the cross product $\mathbf{B} \times \mathbf{C}$ equals the area of the parallelogram formed by vectors $\mathbf{B}$ and $\mathbf{C}$ (Fig. A1.6). Hence $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$ equals the volume of the parallelepiped formed by the three vectors in Fig. A1.6, because $A \cos \beta$ is the height of the parallelepiped. It can be demonstrated that the volume of the same parallelepiped is obtained by the other two triple scalar products in Eq. (A1.10).

## A1.2.7   THE TRIPLE VECTOR PRODUCT

The other product of three vectors has the form $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ and is known as a *triple vector product*. This is a vector normal to both vectors $\mathbf{A}$ and $\mathbf{B} \times \mathbf{C}$. Since the vector $\mathbf{B} \times \mathbf{C}$ is normal to the plane of vectors $\mathbf{B}$ and $\mathbf{C}$, it follows that the vector $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ lies in the plane of vectors $\mathbf{B}$ and $\mathbf{C}$. Therefore, $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = a\mathbf{B} + b\mathbf{C}$, where $a$ and $b$ are scalars to be expressed in terms of vectors $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. It can be proved (e.g., by expanding the left-hand side and the right-hand side in the rectangular coordinate system) that

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C}) \, \mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C} = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}. \tag{A1.11}$$

*Questions and problems:*   QA1.3 to QA1.13, PA1.1 to PA1.6

# A1.3   Orthogonal Coordinate Systems

When vectors are used in actual calculations, they are usually represented in a convenient coordinate system. Generally speaking, a *coordinate system* is a structure that enables unique definition of a point in space by three variables, known as the coordinates of the point. The lines along which two of the three variables are constant are

known as coordinate lines. If coordinate lines are mutually orthogonal at all points of space, the coordinate system is termed an *orthogonal coordinate system*. We describe three orthogonal coordinate systems most often needed in electromagnetics: the rectangular, the cylindrical, and the spherical. The rectangular system is useful for a rectangular waveguide cavity, for example. The cylindrical system is needed, for example, for a coaxial cable. The spherical system is a natural one to use when dealing with antennas, which radiate in all radial directions in space.

### A1.3.1    RECTANGULAR COORDINATE SYSTEM

The rectangular, or cartesian, coordinate system (Fig. A1.7a) is the most commonly used. A point in the rectangular system is defined by coordinates $(x, y, z)$, which represent distances of the point from the planes $x = 0$, $y = 0$, and $z = 0$. A point can also be defined as an intersection of three *coordinate surfaces*—in this case planes (Fig. A1.7b).

The directions of the three axes in a rectangular coordinate system are characterized by three mutually orthogonal unit vectors, $\mathbf{u}_x$, $\mathbf{u}_y$, $\mathbf{u}_z$, known as the *base unit vectors*, shown in Fig. A1.8a. Note that they have the same direction at all points, which is the case only for the rectangular system. A vector represented by its three rectangular components is shown in Fig. A1.8b. The magnitudes of these components, $A_x$, $A_y$, and $A_z$, are known as the *scalar components* of the vector.

Note the sequence of the base unit vectors. It is such that if the first rotates toward the second following the smaller angle between them, the direction of the third is obtained by the right-hand rule. For this reason the rectangular coordinate system is known as a *right-handed coordinate system*. The cylindrical and spherical coordinate systems are also defined so as to be right-handed systems.

We will need the differential lengths $dl_1$, $dl_2$, and $dl_3$ corresponding to a small increase in *one* coordinate, with the other two kept constant. In a rectangular coordinate system these differential lengths (Fig. A1.9) evidently are



(a)                                    (b)
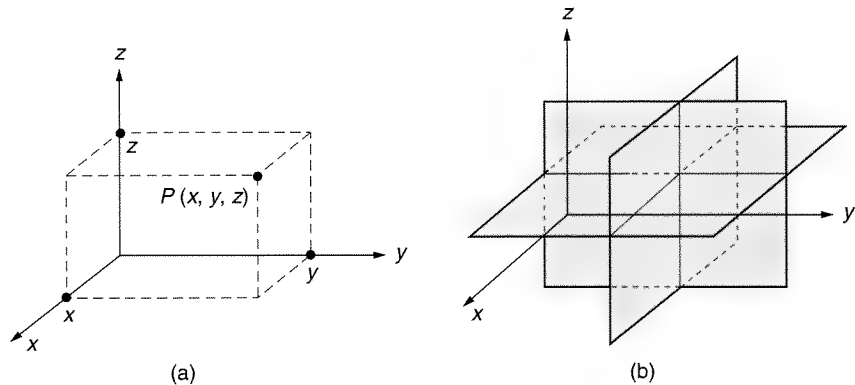
Figure A1.7 (a) Rectangular coordinates of a point. (b) The definition of a point in a rectangular coordinate system can also be an intersection of three coordinate surfaces.

(a)                                              (b)
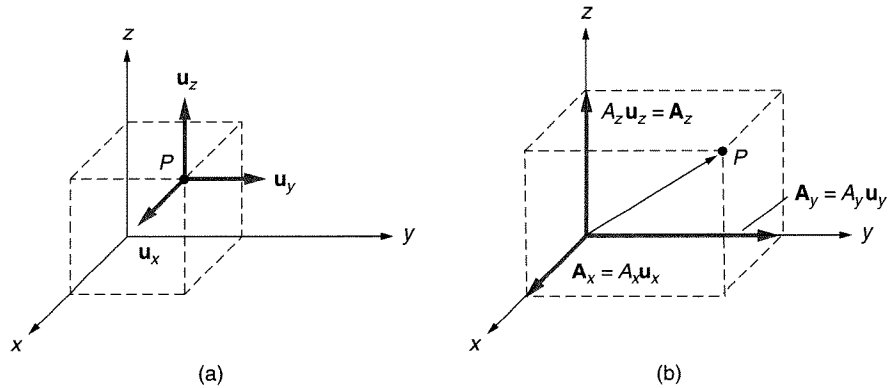
Figure A1.8 (a) The base unit vectors in a rectangular coordinate system. (b) A vector with a starting point $P$ decomposed into its three rectangular components.

THE END

$$dl_1 = dx, \qquad dl_2 = dy, \qquad dl_3 = dz. \qquad\qquad (A1.12)$$

*(Differential lengths along coordinate lines in rectangular coordinate system)*

**Example A1.5—Dot product in the rectangular coordinate system.** In the rectangular coordinate system, vectors **A** and **B** are represented in terms of their scalar components as

$$\mathbf{A} = A_x\mathbf{u}_x + A_y\mathbf{u}_y + A_z\mathbf{u}_z$$

and

$$\mathbf{B} = B_x\mathbf{u}_x + B_y\mathbf{u}_y + B_z\mathbf{u}_z.$$
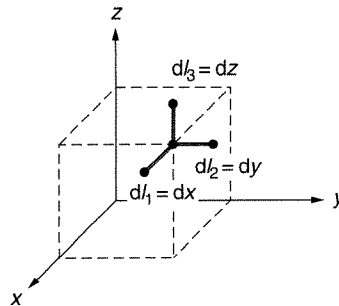


Figure A1.9 Illustration of the differential lengths along coordinate lines in a rectangular coordinate system

Since the base unit vectors are mutually orthogonal, the dot product of any two is zero, and the dot product of any of them by itself is unity. So we obtain

$$\mathbf{A} \cdot \mathbf{B} = A_x B_x + A_y B_y + A_z B_z. \tag{A1.13}$$

*(Dot product in the rectangular coordinate system)*

**Example A1.6—Cross product in the rectangular coordinate system.** With the vectors **A** and **B** represented as in Example A1.5, we have

$$\mathbf{A} \times \mathbf{B} = (A_x \mathbf{u}_x + A_y \mathbf{u}_y + A_z \mathbf{u}_z) \times (B_x \mathbf{u}_x + B_y \mathbf{u}_y + B_z \mathbf{u}_z).$$

Note that the cross product of any unit vector with itself is zero (because the angle is zero), and that

$$\mathbf{u}_x \times \mathbf{u}_y = -\mathbf{u}_y \times \mathbf{u}_x = \mathbf{u}_z \qquad \mathbf{u}_x \times \mathbf{u}_z = -\mathbf{u}_z \times \mathbf{u}_x = -\mathbf{u}_y \qquad \mathbf{u}_y \times \mathbf{u}_z = -\mathbf{u}_z \times \mathbf{u}_y = \mathbf{u}_x,$$

so that

$$\mathbf{A} \times \mathbf{B} = \mathbf{u}_x (A_y B_z - A_z B_y) + \mathbf{u}_y (A_z B_x - A_x B_z) + \mathbf{u}_z (A_x B_y - A_y B_x). \tag{A1.14}$$

*(Cross product in the rectangular coordinate system)*

### A1.3.2  CYLINDRICAL COORDINATE SYSTEM

In the cylindrical coordinate system (Fig. A1.10a), the coordinates of a point are the distance $r$ from the coordinate origin of the projection of the point onto the plane $z = 0$, the angle $\phi$ made by this projection with the $x$ axis, and the $z$ coordinate as in
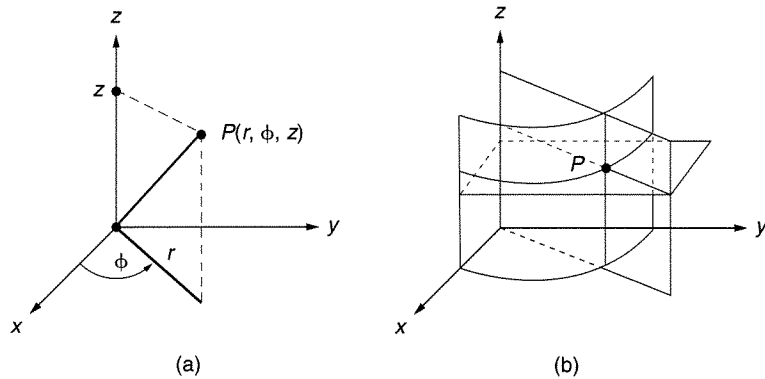


Figure A1.10 (a) Cylindrical coordinates of a point. (b) The definition of a point in the cylindrical coordinate system as the intersection of three coordinate surfaces.
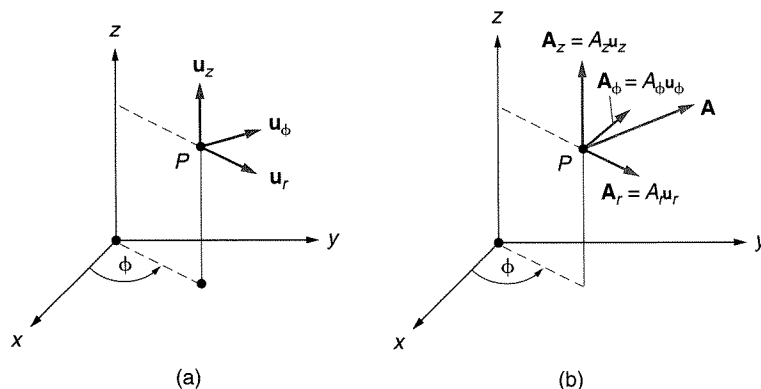
Figure A1.11 (a) The base unit vectors in a cylindrical coordinate system. (b) A vector with a starting point $P$ decomposed into its three cylindrical components.

the rectangular system. The coordinate surfaces (whose intersection defines a point) are circular cylinders of radius $r$, half-planes $\phi = $ constant, and planes $z = $ constant (Fig. A1.10b).

The base unit vectors are in the directions of motion *if* the coordinate of interest is given a small positive increment and the other two are kept fixed. Thus the base unit vectors $\mathbf{u}_r$, $\mathbf{u}_\phi$, and $\mathbf{u}_z$ are as in Fig. A1.11a. Note that although the base unit vector $\mathbf{u}_z$ has the same direction at all points, the base unit vectors $\mathbf{u}_r$ and $\mathbf{u}_\phi$ generally differ from one point to another. Note that the three unit vectors are mutually orthogonal. A vector represented by its three cylindrical components is shown in Fig. A1.11b.

Note that the order of the base unit vectors is again such that if the first rotates toward the second following the smaller angle between them, the direction of the third is obtained by the right-hand rule. Therefore the cylindrical coordinate system is also an orthogonal right-handed system.

The differential lengths $dl_1$, $dl_2$, and $dl_3$ along coordinate lines corresponding to a small increase in *one* coordinate with the other two kept constant (Fig. A1.12) have the form

$$dl_1 = dr, \qquad dl_2 = r\,d\phi, \qquad dl_3 = dz. \tag{A1.15}$$

*(Differential lengths along coordinate lines in the cylindrical coordinate system)*

## A1.3.3  SPHERICAL COORDINATE SYSTEM

The last coordinate system we need is the spherical one. In terms of rectangular coordinates, a point in the spherical system is defined by its distance $r$ from the origin, the angle $\theta$ that $r$ makes with the $z$ axis, and the angle $\phi$ that the projection of $r$ onto
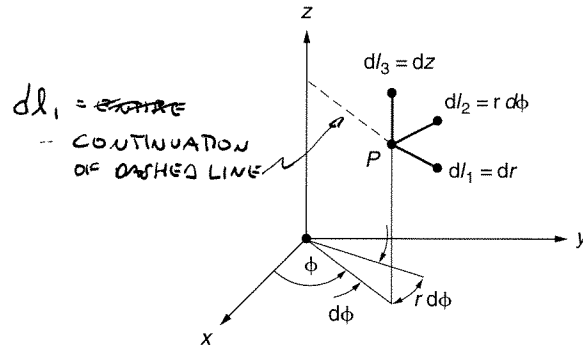
Figure A1.12 Differential lengths in a cylindrical coordinate system

the plane $z = 0$ makes with the $x$ axis (Fig. A1.13a). A point can also be considered as an intersection of three coordinate surfaces, a sphere of radius $r$, a circular cone of half-angle $\theta$, and a half plane containing the $z$ axis and making an angle $\phi$ with the $x$ axis, as in Fig. A1.13b.

The directions of the base unit vectors at a point in a spherical coordinate system are obtained if one coordinate at a time is given as a small positive increment, the other two being kept constant (Fig. A1.14a). Note that they are mutually orthogonal and that the direction of *all three* varies from point to point. A vector decomposed into its three spherical components is shown in Fig. A1.14b. The sequence of the base unit vectors $(\boldsymbol{u}_r, \boldsymbol{u}_\theta, \boldsymbol{u}_\phi)$ is again such that the defined spherical coordinate system is a right-handed system.
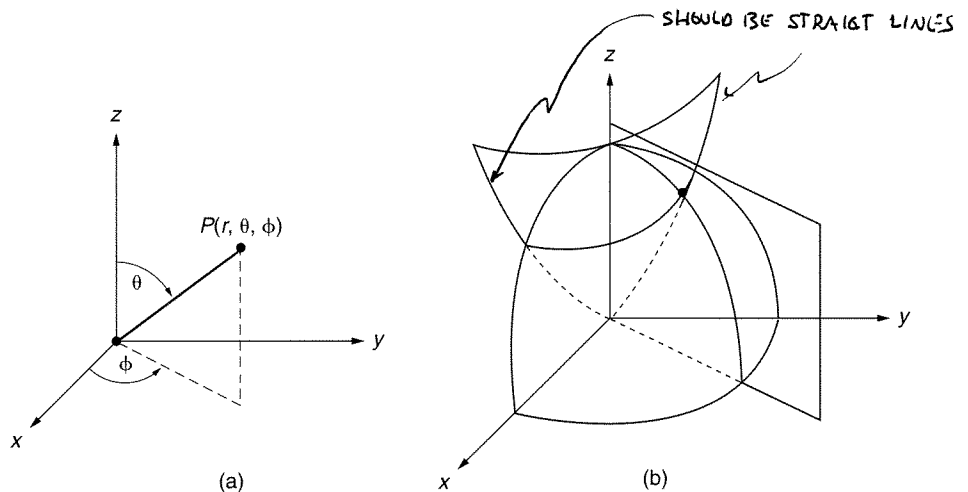


Figure A1.13 (a) Spherical coordinates of a point. (b) The definition of a point in spherical coordinate system as the intersection of three coordinate surfaces.
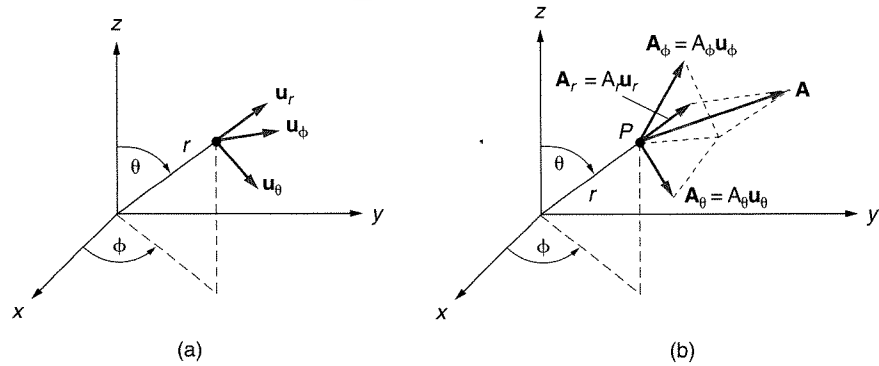
Figure A1.14 (a) The base unit vectors in a spherical coordinate system. (b) A vector with a starting point $P$ decomposed into its three spherical components.
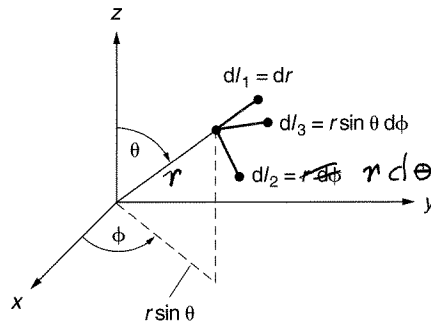


Figure A1.15 Differential lengths along coordinate lines in a spherical coordinate system

The differential lengths $dl_1$, $dl_2$, and $dl_3$ along the three coordinate lines in a spherical coordinate system (Fig. A1.15) are

$$dl_1 = d, \qquad dl_2 = r \, d\theta, \qquad dl_3 = r \sin\theta \, d\phi. \tag{A1.16}$$

*(Differential lengths along coordinate lines in the spherical coordinate system)*

*Questions and problems:*    QA1.14 to QA1.27, PA1.7 to PA1.24

## A1.4   Elements of Vector Calculus

Vector calculus deals with various forms of differentiations and integrations of vector functions. The bases for most of the conclusions are the definitions of three types of

"derivatives" (actually, a combination of derivatives with respect to the three coordinates) that can be associated with scalar and vector fields. The first of these is the gradient. It operates on a scalar field and yields a vector field. The second is the divergence. It operates on a vector field, resulting in a scalar field. The third is the curl. It also operates on a vector field, but the result is a new vector field.

These basic operations can be combined to give second-order operations such as divergence of a gradient, divergence of a curl, curl of a gradient, and gradient of a scalar product of two vector fields.

## A1.4.1  GRADIENT

The values of the function describing a scalar field in the neighborhood of a point are generally not the same. If we move in various directions from the point considered, in some directions the values of the function gradually decrease, and in some they gradually increase. The *gradient* is a *vector* function of coordinates pointing in the direction of the *most rapid increase* of these gradual variations (hence the name) of the scalar field in the neighborhood of a point. For example, a temperature inside a room generally differs from one point to another. The gradient of temperature at a point will be directed toward the warmest neighboring points.

Let a scalar field be described by the function $f(c_1, c_2, c_3)$, where ($c_1$, $c_2$, and $c_3$) are the coordinates in any coordinate system (rectangular, cylindrical, spherical, or any other). Consider a point $A$ in the field and assume that an axis $\zeta$ passes through that point, as in Fig. A1.16. Let the value of the function at $A$ be $f$, and at a very close point $B$ on the $\zeta$ axis let a distance $d\zeta$ from the point $A$ in the reference axis direction be $f + df_\zeta$. The expression

$$\frac{(f + df_\zeta) - f}{d\zeta} = \frac{df_\zeta}{d\zeta}$$

is evaluated as the ordinary derivative of a function of a single unknown, $\zeta$. Becuase it is associated with a specific direction (that of the $\zeta$ axis), it is known as the *directional derivative*. If we multiply this expression by the unit vector $\mathbf{u}_\zeta$ along the $\zeta$ axis, we obtain a vector with the following properties: (1) the magnitude of the vector equals the directional derivative of $f$ in the $\zeta$ axis direction, and (2) the direction of the vector coincides with that of the $\zeta$ axis. Thus, instead of giving the value of the directional derivative and of explaining in which direction it was obtained, we specify
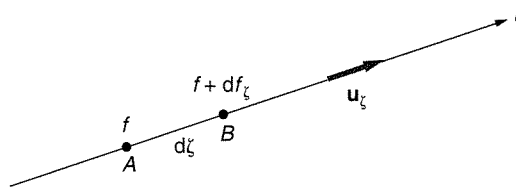


Figure A1.16 A $\zeta$ axis in a scalar field described by a function $f$

just one vector, which contains all the information we need. This vector is known as the *component of the gradient of the function f in the direction of the ζ axis* and is denoted as $\text{grad}_\zeta f = \nabla_\zeta f$:

$$\nabla_\zeta f = \frac{df_\zeta}{d\zeta} \mathbf{u}_\zeta. \tag{A1.17}$$

*(Component of the gradient in the direction of a ζ axis)*

This procedure can be used to obtain the component of the gradient in any direction, and thus also along the three local coordinate axes defined by the base unit vectors. Of course, the derivatives will be partial derivatives, because in determining the derivative in Eq. (A1.17) in such a case we change only one coordinate at a time. The total vector $\nabla f$, as any vector, is obtained as a sum of these three components. Note that in Eq. (A1.17) the ζ coordinate is assumed to be a *length coordinate*, so that the increment $d\zeta$ is also a differential length.

That the vector $\nabla f$ is in the direction of the maximal increase of the function $f$ at a point is almost obvious: if we make a vector sum of the rate of increase of $f$ in three orthogonal directions, the result will be the direction of the maximal increase of $f$ in space.

We know differential length elements in rectangular, cylindrical, and spherical coordinate systems. In a rectangular coordinate system, from Eqs. (A1.17) and (A1.12) we obtain the following expression for the gradient of the function $f(x, y, z)$:

$$\nabla f(x, y, z) = \frac{\partial f(x, y, z)}{\partial x} \mathbf{u}_x + \frac{\partial f(x, y, z)}{\partial y} \mathbf{u}_y + \frac{\partial f(x, y, z)}{\partial z} \mathbf{u}_z. \tag{A1.18}$$

*(Gradient in the rectangular coordinate system)*

Note that the *del operator*, or *nabla operator*, is defined in the rectangular system as

$$\nabla = \frac{\partial}{\partial x} \mathbf{u}_x + \frac{\partial}{\partial y} \mathbf{u}_y + \frac{\partial}{\partial z} \mathbf{u}_z. \tag{A1.19}$$

*(Definition of del operator)*

Although defined in a rectangular coordinate system, the symbol $\nabla f$ is used frequently to denote the gradient in *any* coordinate system.

The differential lengths along coordinate lines in the cylindrical and spherical coordinate systems are given in Eqs. (A1.15) and (A1.16). From Eq. (A1.17) the ex-

pressions for the gradient in the two systems are

$$\nabla f(r, \phi, z) = \frac{\partial f(r, \phi, z)}{\partial r} \mathbf{u}_r + \frac{1}{r} \frac{\partial f(r, \phi, z)}{\partial \phi} \mathbf{u}_\phi + \frac{\partial f(r, \phi, z)}{\partial z} \mathbf{u}_z. \qquad (A1.20)$$
*(Gradient in the cylindrical coordinate system)*

and

$$\nabla f(r, \theta, \phi) = \frac{\partial f(r, \theta, \phi)}{\partial r} \mathbf{u}_r + \frac{1}{r} \frac{\partial f(r, \theta, \phi)}{\partial \theta} \mathbf{u}_\theta + \frac{1}{r \sin \theta} \frac{\partial f(r, \theta, \phi)}{\partial \phi} \mathbf{u}_\phi. \qquad (A1.21)$$
*(Gradient in the spherical coordinate system)*

## A1.4.2  DIVERGENCE

The *divergence* of a vector field is a *scalar* function indicating integrally how much the field "diverges," i.e., originates, from a small closed surface enclosing a point in the field. For example, if we imagine a point in air from which there is a permanent outflow of air, the divergence at that point would be a positive number. If there is an inflow of air at the point, the divergence would be a negative number. At points with no outflow or inflow, the divergence is zero. The divergence of a vector function $\mathbf{F}$ is denoted as $\nabla \cdot \mathbf{F}$.

To represent this explanation mathematically, consider a point in the vector field described by a vector function $\mathbf{F}(c_1, c_2, c_3)$ of three arbitrary orthogonal coordinates $c_1$, $c_2$, and $c_3$. Let $\Delta S$ be a small surface enclosing the point, and let $\Delta v$ be the volume enclosed by $\Delta S$. The divergence of $\mathbf{F}(c_1, c_2, c_3)$ at that point is then defined as

$$\nabla \cdot \mathbf{F}(c_1, c_2, c_3) = \lim_{\Delta v \to 0} \frac{1}{\Delta v} \oint_{\Delta S} \mathbf{F} \cdot d\mathbf{S}, \qquad (A1.22)$$
*(Definition of divergence of a function $\mathbf{F}$)*

The product $\mathbf{F} \cdot d\mathbf{S}$ is known as the *flux* of the vector $\mathbf{F}$ through the differential surface $d\mathbf{S}$. Note that only the component of $\mathbf{F}$ *normal* to $d\mathbf{S}$ contributes to the flux.

If we have a differential volume in the form of a parallelepiped (box) defined by three differential length elements in a coordinate system, we have to integrate over its six sides. Pairs of opposite parallelepiped surfaces are normal to one of the three base unit vectors. Therefore the flux through such a pair will be due to only one component of the vector. For example, the flux through the sides 1 and 1' in Fig. A1.17 is due only to the component $\mathbf{F}_1$ of the vector function $\mathbf{F}$.
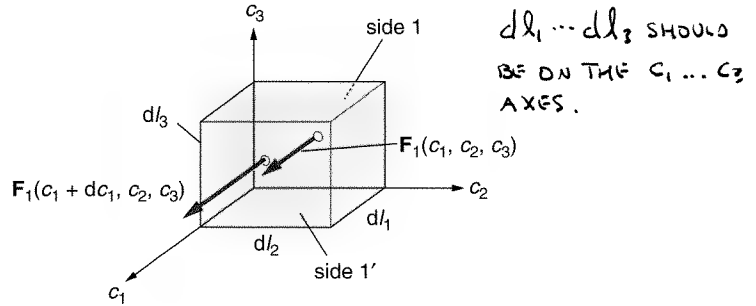
Figure A1.17 A pair of differential surfaces corresponding to fixed values of the second and third coordinate, $c_2$ and $c_3$, for two values, $c_1$ and $c_1 + dc_1$, of the first coordinate

There is only one problem that remains to be solved. How can we determine the flux of the vector function $F_1(c_1, c_2, c_3)$ when it is given as a *general* function of the coordinates? Let this be the value of the function on side 1 in Fig. A1.17. The flux through side 1 (note that it is an inward flux, i.e., negative) is then

Flux through side 1 $= -F_1(c_1, c_2, c_3) \, dl_2 \, dl_3$.

To obtain the outward, i.e., positive flux through side 1′, we need to consider the *flux through side 1*, and not only the function $F_1(c_1, c_2, c_3)$, as a function of the coordinate $c_1$. This is necessary because the size of the side 1′ is also a function of $c_1$. So

Flux through side 1′ $= F_1(c_1, c_2, c_3) \, dl_2 \, dl_3 + \dfrac{\partial[F_1(c_1, c_2, c_3) \, dl_2 \, dl_3]}{\partial c_1} \, dc_1$.

Adding the two fluxes, we obtain that the flux through the pair of surfaces 1 and 1′ in Fig. A1.17 is

Flux through sides 1 and 1′ $= \dfrac{\partial[F_1(c_1, c_2, c_3) \, dl_2 \, dl_3]}{\partial c_1} \, dc_1$

The flux through the other two pairs is obtained in exactly the same way. When we sum the three fluxes and divide by the differential volume, $dl_1 \, dl_2 \, dl_3$, we obtain a general expression for the divergence:

$$\nabla \cdot \mathbf{F} = \frac{1}{dl_1 \, dl_2 \, dl_3} \left[ \frac{\partial(F_1 \, dl_2 \, dl_3)}{\partial c_1} \, dc_1 + \frac{\partial(F_2 \, dl_1 \, dl_3)}{\partial c_2} \, dc_2 + \frac{\partial(F_3 \, dl_1 \, dl_2)}{\partial c_3} \, dc_3 \right],$$

(A1.23)

*(Divergence in orthogonal coordinates $c_1$, $c_2$, and $c_3$)*

where

$\mathbf{F} = \mathbf{F}(c_1, c_2, c_3) = F_1(c_1, c_2, c_3)\mathbf{u}_1 + F_2(c_1, c_2, c_3)\mathbf{u}_2 + F_3(c_1, c_2, c_3)\mathbf{u}_3$.

Using this formula and the expressions for differential lengths in rectangular, cylindrical, and spherical coordinate systems and noting that the coordinates are independent variables, we obtain directly the expressions for the divergence in rectangular, cylindrical, and spherical coordinate systems. In a rectangular system $c_1 = x$, $c_2 = y$, $c_3 = z$, so that

$$\nabla \cdot \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_x}{\partial y} + \frac{\partial F_x}{\partial z}. \tag{A1.24}$$

*(Divergence in a rectangular coordinate system)*

In a cylindrical system $c_1 = r$, $c_2 = \phi$, $c_3 = z$, and the expression in Eq. (A1.23) becomes

$$\nabla \cdot \mathbf{F} = \frac{1}{r} \frac{\partial (r F_r)}{\partial r} + \frac{1}{r} \frac{\partial F_\phi}{\partial \phi} + \frac{\partial F_z}{\partial z}. \tag{A1.25}$$

*(Divergence in a cylindrical coordinate system)*

In a spherical system $c_1 = r$, $c_2 = \theta$, $c_3 = \phi$, so that Eq. (A1.23) takes the form

$$\nabla \cdot \mathbf{F} = \frac{1}{r^2} \frac{\partial (r^2 F_r)}{\partial r} + \frac{1}{r \sin \theta} \frac{\partial (\sin \theta \, F_\theta)}{\partial \theta} + \frac{1}{r \sin \theta} \frac{\partial F_\phi}{\partial \phi}. \tag{A1.26}$$

*(Divergence in a spherical coordinate system)*

## A1.4.3   CURL

In the velocity vector field of air, for example, you will notice that a light feather is not only being translated but usually also rotates. This is due to different air velocities normal to the main motion of air, resulting in miniature whirlwinds all over the vector field. This property of any vector field (not necessarily describing a motion), of having different magnitude at close points transverse to the field lines, is described by a vector quantity known as the *curl* of the vector field. It is a vector normal to the plane in which this curling action of the field at a point is maximal, and is denoted as $\nabla \times \mathbf{F}$.

As a vector quantity, the curl can be obtained as a sum of its three components. So we define first the component of the curl of a vector function $\mathbf{F}$ in the direction of a unit vector $\mathbf{n}$, as in Fig. A1.18. Assume a small surface of area $\Delta S$ bounded by a contour $\Delta C$, *normal* to $\mathbf{n}$. Let the direction around the contour be connected with the direction of the unit vector $\mathbf{n}$ by the right-hand screw rule. The component of $\nabla \times \mathbf{F}$ in the direction of the unit vector $\mathbf{n}$, that is, $\mathbf{n} \cdot \nabla \times \mathbf{F}$, is then defined as

$$\mathbf{n} \cdot \nabla \times \mathbf{F} = \lim_{\Delta S \to 0} \frac{1}{\Delta S} \oint_{\Delta C} \mathbf{F} \cdot d\mathbf{l}. \tag{A1.27}$$

*[Component of curl of vector $\mathbf{F}$ in the direction of unit vector $\mathbf{n}$ (Fig. A1.18)]*
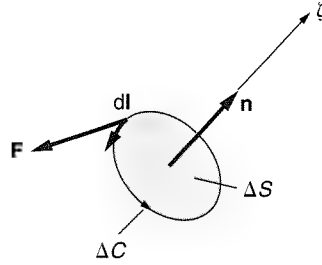
Figure A1.18 A surface $\Delta S$ bounded by the contour $\Delta C$ normal to unit vector $\mathbf{n}$

Using this definition, it is not difficult to obtain a general expression for the curl in orthogonal coordinate systems. Again let the vector field be described by a vector function $\mathbf{F}(c_1, c_2, c_3)$, where $(c_1, c_2, c_3)$ are any orthogonal coordinates. Let us find the expression for the $c_1$ component of $\nabla \times \mathbf{F}$, Fig. A1.19.

In calculating the integral around the rectangular contour $\Delta C$ in Fig. A1.19, we have four terms. Since the dot product is integrated, we need only components of vector $\mathbf{F}$ along the rectangle sides. Omitting for simplicity the dependence on coordinates $(c_1, c_2, c_3)$, the contribution to the integral around $\Delta C$ along sides 2 and 2' is

$$\int_{\text{side 2}} \mathbf{F} \cdot d\mathbf{l} + \int_{\text{side 2'}} \mathbf{F} \cdot d\mathbf{l} = F_2 \, dl_2 - \left[ F_2 \, dl_2 + \frac{\partial(F_2 \, dl_2)}{\partial c_3} \, dc_3 \right] = \frac{-\partial(F_2 \, dl_2)}{\partial c_3} \, dc_3.$$

Similarly, the contribution to the integral along sides 3 and 3' is

$$\int_{\text{side 3}} \mathbf{F} \cdot d\mathbf{l} + \int_{\text{side 3'}} \mathbf{F} \cdot d\mathbf{l} = -F_3 \, dl_3 + \left[ F_3 \, dl_3 + \frac{\partial(F_3 \, dl_3)}{\partial c_2} \, dc_2 \right] = \frac{\partial(F_3 \, dl_3)}{\partial c_2} \, dc_3.$$

Thus, the area of the rectangle being $dl_2 \, dl_3$, according to the definition in Eq. (A1.27) the $c_1$ component of $\nabla \times \mathbf{F}$ is

$$\mathbf{u}_1 \cdot \nabla \times \mathbf{F} = \frac{1}{dl_2 \, dl_3} \left[ \frac{\partial(F_3 \, dl_3)}{\partial c_2} \, dc_2 - \frac{\partial(F_2 \, dl_2)}{\partial c_3} \, dc_3 \right].$$

The $c_2$ and $c_3$ components of the curl are obtained in the same manner, so that the final general expression for the curl of vector $\mathbf{F}$ in any orthogonal coordinate
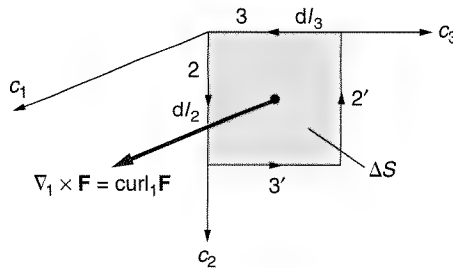
Figure A1.19 Determination of the $c_1$ component of $\nabla \times \mathbf{F}$

system is

$$
\nabla \times \mathbf{F} = \mathbf{u}_1 \frac{1}{dl_2 \, dl_3} \left[ \frac{\partial (F_3 \, dl_3)}{\partial c_2} \, dc_2 - \frac{\partial (F_2 \, dl_2)}{\partial c_3} \, dc_3 \right]
$$
$$
+ \mathbf{u}_2 \frac{1}{dl_1 \, dl_3} \left[ \frac{\partial (F_1 \, dl_1)}{\partial c_3} \, dc_3 - \frac{\partial (F_3 \, dl_3)}{\partial c_1} \, dc_1 \right]
$$
$$
+ \mathbf{u}_3 \frac{1}{dl_1 \, dl_2} \left[ \frac{\partial (F_2 \, dl_2)}{\partial c_1} \, dc_1 - \frac{\partial (F_1 \, dl_1)}{\partial c_2} \, dc_2 \right]. \tag{A1.28}
$$

*(Curl in orthogonal coordinates $c_1$, $c_2$, and $c_3$)*

Because we know the length elements in rectangular, cylindrical, and spherical coordinate systems, using this general formula we easily find the expression for the curl in the three systems. In a rectangular system, where $c_1 = x$, $c_2 = y$, and $c_3 = z$, we obtain

$$
\nabla \times \mathbf{F} = \mathbf{u}_x \left( \frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right) + \mathbf{u}_y \left( \frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x} \right) + \mathbf{u}_z \left( \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right). \tag{A1.29}
$$

*(Curl in a rectangular coordinate system)*

In a cylindrical system, $c_1 = r$, $c_2 = \phi$, $c_3 = z$, and the expression in Eq. (A1.28) results in

$$
\nabla \times F = \mathbf{u}_r \left( \frac{1}{r} \frac{\partial F_z}{\partial \phi} - \frac{\partial F_\phi}{\partial z} \right) + \mathbf{u}_\phi \left( \frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r} \right) + \mathbf{u}_z \frac{1}{r} \left[ \frac{\partial (r F_\phi)}{\partial r} - \frac{\partial F_r}{\partial \phi} \right]. \tag{A1.30}
$$

*(Curl in a cylindrical coordinate system)*

In a spherical system, $c_1 = r$, $c_2 = \theta$, $c_3 = \phi$, so that according to Eq. (A1.28) the curl becomes

$$
\nabla \times F = \mathbf{u}_r \frac{1}{r \sin \theta} \left[ \frac{\partial (\sin \theta \, F_\phi)}{\partial \theta} - \frac{\partial F_\theta}{\partial \phi} \right]
$$
$$
+ \mathbf{u}_\theta \frac{1}{r} \left[ \frac{1}{\sin \theta} \frac{\partial F_r}{\partial \phi} - \frac{\partial (r F_\phi)}{\partial r} \right] + \mathbf{u}_\phi \frac{1}{r} \left[ \frac{\partial (r F_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right]. \tag{A1.31}
$$

*(Curl in a spherical coordinate system)*

**Example A1.7—Divergence of the cross product of two vector functions.** Consider the expression $\nabla \cdot (\mathbf{A} \times \mathbf{B})$, where $\mathbf{A}$ and $\mathbf{B}$ are vector functions. This is a legitimate operation, since the cross product of two vectors is again a vector. Because the del operator is a differential operator, we have to treat this expression as a derivative of a product. So we have

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \nabla \cdot (\mathbf{A} \times [\mathbf{B}]) + \nabla \cdot ([\mathbf{A}] \times \mathbf{B}),$$

where the brackets indicate that the vector function in the particular operation should be considered as a constant.

The del operator is a vector operator, so formally we can make use of the property of the triple scalar product in Eq. (A1.10), to obtain

$$\nabla \cdot (\mathbf{A} \times [\mathbf{B}]) = [\mathbf{B}] \cdot (\nabla \times \mathbf{A}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}), \tag{A1.32}$$

and

$$\nabla \cdot ([\mathbf{A}] \times \mathbf{B}) = [\mathbf{A}] \cdot (\mathbf{B} \times \nabla) = -\mathbf{A} \cdot (\nabla \times \mathbf{B}).$$

So we finally have

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B}). \tag{A1.33}$$

## A1.4.4  LAPLACIAN OF A SCALAR FIELD

Because the gradient is a vector, a legitimate operation is to find the divergence of that vector, i.e., to find $\nabla \cdot (\nabla V)$. This is known as the *Laplacian of V*. Since the gradient implies differentiation and so does the divergence, a combination of second derivatives of the scalar function $V$ results from this operation.

The symbol for the Laplacian of $V$ is denoted by $\nabla^2 V$ (symbolizing the operation $\nabla \cdot (\nabla V)$, or sometimes $\Delta V$. The symbol $\nabla^2 = \nabla \cdot \nabla$ is known as the *Laplacian operator*. In the rectangular coordinate system it has the form

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{A1.34}$$

*(Laplacian operator in rectangular coordinates)*

Because we know the gradient and the divergence in rectangular, cylindrical, and spherical coordinate systems, it is not difficult to determine the Laplacian of a scalar function in these systems. For example, in the rectangular system, where we know the Laplacian operator,

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2}. \tag{A1.35}$$

*(Laplacian of a scalar function in a rectangular coordinate system)*

### A1.4.5  LAPLACIAN OF A VECTOR FIELD

Curl of a curl of a vector function is a legitimate operation because curl of a vector is also a vector. It is again a second-order operation, resulting in a combination of second derivatives of the components of the vector function. It is written as $\nabla \times (\nabla \times \mathbf{F})$. Although it is an operator, $\nabla$ can be considered as a vector, so that the last expression can be expanded using the identity in Eq. (A1.11), to obtain

$$\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - (\nabla \cdot \nabla)\mathbf{F}, \tag{A1.36}$$

from which, recalling that $\nabla \cdot \nabla = \nabla^2$,

$$\nabla^2 \mathbf{F} = \nabla(\nabla \cdot \mathbf{F}) - \nabla \times (\nabla \times \mathbf{F}). \tag{A1.37}$$

*(Definition of the Laplacian of a vector function)*

The expression on the left-hand side of this equation is known as the *Laplacian of the vector function* $\mathbf{F}$, and the right-hand side tells us how it can be evaluated in any coordinate system. In the rectangular system it can be obtained directly from its definition:

$$\nabla^2 \mathbf{F} = \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) (F_x \mathbf{u}_x + F_y \mathbf{u}_y + F_z \mathbf{u}_z).$$

After performing the indicated formal multiplications, this becomes

$$\nabla^2 \mathbf{F} = \left( \frac{\partial^2 F_x}{\partial x^2} + \frac{\partial^2 F_x}{\partial y^2} + \frac{\partial^2 F_x}{\partial z^2} \right) \mathbf{u}_x$$
$$+ \left( \frac{\partial^2 F_y}{\partial x^2} + \frac{\partial^2 F_y}{\partial y^2} + \frac{\partial^2 F_y}{\partial z^2} \right) \mathbf{u}_x + \left( \frac{\partial^2 F_z}{\partial x^2} + \frac{\partial^2 F_z}{\partial y^2} + \frac{\partial^2 F_z}{\partial z^2} \right) \mathbf{u}_x. \tag{A1.38}$$

Note that $F_x$, $F_y$, and $F_z$ are *scalar functions* (these are scalar components of the vector $\mathbf{F}$). Therefore the expressions in the parentheses in the last equation are the Laplacians of the three scalar components of the vector $\mathbf{F}$. Thus we finally obtain

$$\nabla^2 \mathbf{F} = (\nabla^2 F_x)\mathbf{u}_x + (\nabla^2 F_y)\mathbf{u}_y + (\nabla^2 F_z)\mathbf{u}_z. \tag{A1.39}$$

*(Laplacian of vector **F** in rectangular coordinate system)*

## A1.4.6  DIVERGENCE THEOREM

The divergence theorem is a direct consequence of the definition of the divergence in Eq. (A1.22). It states that the volume integral of the divergence of a vector function **F** over any volume equals the flux of **F** through the surface $S$ enclosing $v$.

For a small surface $\Delta S$, without going to the limit and multiplying Eq. (A1.22) by $\Delta v$, we obtain

$$\nabla \cdot \mathbf{F}\,\Delta v = \oint_{\Delta S} \mathbf{F} \cdot d\mathbf{S}.$$

Imagine now a large domain of volume $v$, enclosed by a large surface $S$, subdivided into small volumes $\Delta v$ enclosed by small surfaces $\Delta S$. For all these small volumes we can write the preceding equation. Adding these equations, the left-hand side becomes the integral over volume $v$ of $\nabla \cdot F\,dv$. The right-hand side represents the sum of the fluxes through all the small closed surfaces pressed one onto another and enclosed by the large surface $S$.

The flux is evaluated always with respect to the *outward* normal, as indicated in Fig. A1.20. The figure shows the cross section of a few small surfaces in the vicinity of the large surface. The fluxes through shared sides of two small closed surfaces is the same, but of opposite sign. Consequently, the flux through those sides of any small closed surface shared by another small closed surface cancels out. Only the flux through that side of a small surface that is a part of the large surface, $S$, is not canceled out. Therefore the sum of fluxes through all the small surfaces equals the
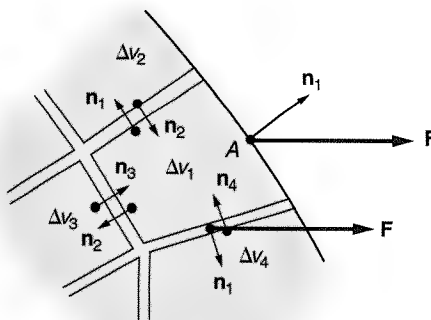


Figure A1.20 Cross section of a volume $v$ consisting of many small subvolumes near the surface $S$ enclosing $v$

flux through the large surface. We obtain an identity, often used in electromagnetics, called the *divergence theorem*:

$$\int_v \nabla \cdot \mathbf{F} \, dv = \oint_S \mathbf{F} \cdot d\mathbf{S}. \qquad\qquad (A1.40)$$

*(Divergence theorem)*

### A1.4.7  STOKES'S THEOREM

Stokes's theorem follows from the definition of the vector component of the curl of a vector function $\mathbf{F}$ in an arbitrary direction, Eq. (A1.29). It states that the flux of $\nabla \times \mathbf{F}$ through any open surface $S$ equals the line integral of $\mathbf{F}$ around the contour bounding $S$.

To prove the theorem, consider a large open surface $S$ and imagine the surface subdivided into many small surface elements $\Delta S$. Figure A1.21 shows one part of $S$ so subdivided, with a few small surface elements near the contour $C$ bounding $S$.

Consider Eq. (A1.27) for one small surface $\Delta S$, but without going to the limit. We multiply this equation by $\Delta S$ to obtain

$$\nabla \times \mathbf{F} \cdot \Delta \mathbf{S} = \oint_{\Delta C} \mathbf{F} \cdot d\mathbf{l};$$

because the product $\Delta S \mathbf{n}$ is the vector surface element, $\Delta \mathbf{S}$.

By adding such equations for all the surface elements, the left-hand side becomes simply the flux of the vector $\overset{\times}{\mathbf{F}}$ through the surface $S$. The line integral along any side of a small contour $\Delta C$ shared by an adjacent small contour cancels out (Fig. A1.21). We are left with the line integral along segments of the small contours belonging to the large contour $C$ bounding $S$. Thus, writing $d\mathbf{S}$ instead of $\Delta \mathbf{S}$,

$\times\ \nabla \times F$
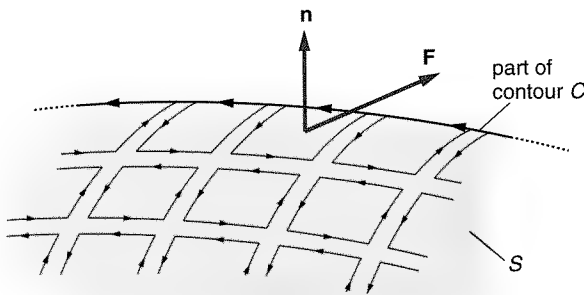


Figure A1.21 Part of the surface $S$ consisting of many small subsurfaces, near the contour $C$ bounding $S$

$$\int_S \nabla \times \mathbf{F} \cdot d\mathbf{S} = \oint_C \mathbf{F} \cdot d\mathbf{l}. \tag{A1.41}$$

*(Stokes's theorem)*

THE DIVERGANCE + Stokes's theorem S HAVE many applications in electromagnetic field theory. We use the two theorems now to derive two important identities of vector calculus.

**Example A1.8—Divergence of the curl is identically zero.** The operation $\nabla \cdot (\nabla \times \mathbf{A})$ is a legitimate operation, since $\nabla \times \mathbf{A}$ is a vector. In Eq. (A1.41) let $\mathbf{F} = \nabla \times \mathbf{A}$. We have

$$\int_v \nabla \cdot (\nabla \times \mathbf{A}) \, dv = \oint_S \nabla \times \mathbf{A} \cdot d\mathbf{S}. \tag{A1.42}$$

The surface integral on the right side will not be changed if we assume the closed surface to have a miniature hole. However, in such a case we can consider the surface to be an *open surface*, with the boundary being the hole boundary. So we can apply Stokes's theorem to the right-hand side. The line integral of $\mathbf{A}$ around the miniature hole tends to zero as the hole shrinks. Therefore we first find that the following integral identity holds:

$$\oint_S \nabla \times \mathbf{A} \cdot d\mathbf{S} \equiv 0. \tag{A1.43}$$

So the left-hand side in Eq. (A1.42) is zero for *any volume*. This is possible only if the integrand is zero, i.e., if

$$\nabla \cdot (\nabla \times \mathbf{A}) = 0. \tag{A1.44}$$

*(Valid for any vector function $\mathbf{A}$)*

**Example A1.9—Curl of the gradient is identically zero.** The operation $\nabla \times (\nabla V)$ is a legitimate operation, since $\nabla V$ is a vector. In Eq. (A1.41) let $F = \mathbf{A} = \nabla V$, to obtain

$$\int_S \nabla \times (\nabla V) \cdot d\mathbf{S} = \oint_C \nabla V \cdot d\mathbf{l}. \tag{A1.45}$$

The integral on the right-hand side can be written in the form

$$\oint_C \nabla V \cdot d\mathbf{l} = \oint_C \frac{\partial V}{\partial l} dl = \oint_C d_l V,$$

where $d_l V$ is the increment of the function $V$ when the point moves from the beginning to the end of the line element $d\mathbf{l}$ of the contour $C$. Since in calculating the integral we start at a point of $C$, describe the full contour, and return to the starting point, the integration amounts to integrating a function from a point to that same point. We know that the result is zero. So the right-hand side in Eq. (A1.45) is zero,

$$\oint_C \nabla V \cdot d\mathbf{l} \equiv 0. \tag{A1.46}$$

For the left-hand side in Eq. (A1.46) to be zero for *any surface S*, the integrand must be identically zero. So we have another identity,

$$\nabla \times (\nabla V) = \nabla \times (\nabla V) = 0. \tag{A1.47}$$

*(Valid for any scalar function V)*

*Questions and problems:*    QA1.28 to QA1.41, PA1.25 to PA1.35

# QUESTIONS

**QA1.1.** Classify the following quantities as either scalar or vector quantities: mass, time, weight, course of a ship, position of a point with respect to another point, acceleration, power of an engine, current intensity, voltage.

**QA1.2.** Classify the following quantities as either scalar or vector fields: temperature in a room, mass density of an inhomogeneous body, weight per unit volume of an inhomogeneous body, velocity of air particles in a room, velocity of water particles in a river.

**QA1.3.** Discuss whether the definition of subtraction of vectors follows from the definition of vector addition, or whether an additional definition is indispensable.

**QA1.4.** Sketch the dependence of the dot product of two unit vectors on the angle between them.

**QA1.5.** Why is the cross product not commutative?

**QA1.6.** Which of the following expressions does not make sense, and why? (1) $\mathbf{A} \times (\mathbf{B} \cdot \mathbf{C})$, (2) $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}$, (3) $(\mathbf{A} \times \mathbf{B}) \times (\mathbf{C} \cdot \mathbf{D})$, (4) $(\mathbf{A} \cdot \mathbf{B})(\mathbf{C} \times \mathbf{D})$, (5) $[(\mathbf{A} \cdot \mathbf{B}) \mathbf{C}] \times \mathbf{D}$, (6) $[(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}] \times \mathbf{D}$, (7) $[\mathbf{A} \times (\mathbf{B} \times \mathbf{C})] \cdot \mathbf{D}$

**QA1.7.** What is the necessary and sufficient condition for three vectors, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, to be in the same plane?

**QA1.8.** Prove that $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{A}$ and $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{B}$ are zero.

**QA1.9.** If $\mathbf{A} \cdot \mathbf{C} = \mathbf{B} \cdot \mathbf{C}$, does it mean that $\mathbf{A} = \mathbf{B}$? Explain.

**QA1.10.** If $\mathbf{A} \times \mathbf{C} = \mathbf{B} \times \mathbf{C}$, does it mean that $\mathbf{A} = \mathbf{B}$? Explain.

**QA1.11.** Can the dot product of two vectors be negative? Can the magnitude of the cross product of two vectors be negative?

**QA1.12.** In which cases is (1) $\mathbf{A} \cdot \mathbf{B} = 0$, (2) $\mathbf{A} \times \mathbf{B} = 0$, and (3) $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = 0$?

**QA1.13.** Explain the meaning of $(\mathbf{A} \cdot \mathbf{B})\mathbf{C}$. Is this the same as $\mathbf{C}(\mathbf{A} \cdot \mathbf{B})$?

**QA1.14.** Which of the following sets of coordinates define a point? (1) $x = 2, y = -4, z = 0$, (2) $r = -4, \phi = 0, z = -1$, (3) $r = 3, \theta = -90°, \phi = 0$

**QA1.15.** How do you obtain the components of a vector $\mathbf{A}$ in the direction of the three base unit vectors in any coordinate system?

**QA1.16.** Define coordinate lines.

**QA1.17.** Define differential lengths along coordinate lines.

**QA1.18.** Define orthogonal coordinate systems.

**QA1.19.** What is a "right-handed coordinate system"?

**QA1.20.** A vector is defined by its three orthogonal components. Determine the vector itself and the unit vector in its direction.

**QA1.21.** A vector is defined by its starting and end points in the rectangular system. Determine its components, the vector itself, and the unit vector in its direction.

**QA1.22.** A point is defined (a) in a cylindrical coordinate system by its coordinates $(r, \phi, z)$, and (b) in a spherical coordinate system by its coordinates $(r, \theta, \phi)$. Find the rectangular coordinates of the point in both cases.

**QA1.23.** A point is defined in a rectangular coordinate system by its coordinates $(x, y, z)$. Find the coordinates of the point in (a) cylindrical and (b) spherical coordinate systems.

**QA1.24.** If $\mathbf{u}_x$, $\mathbf{u}_y$, and $\mathbf{u}_z$ are base unit vectors in the rectangular coordinate system, and $\mathbf{u}_r$ that in the cylindrical coordinate system, what are the values of the following products? (1) $\mathbf{u}_r \cdot \mathbf{u}_x$, (2) $\mathbf{u}_r \cdot \mathbf{u}_y$, (3) $\mathbf{u}_r \cdot \mathbf{u}_z$, (4) $\mathbf{u}_r \times \mathbf{u}_x$, (5) $\mathbf{u}_r \times \mathbf{u}_y$, (6) $\mathbf{u}_r \times \mathbf{u}_z$

**QA1.25.** If $\mathbf{u}_\phi$ is the base unit vector in the cylindrical coordinate system, what are the following products equal to? (1) $\mathbf{u}_\phi \cdot \mathbf{u}_x$, (2) $\mathbf{u}_\phi \cdot \mathbf{u}_y$, (3) $\mathbf{u}_\phi \cdot \mathbf{u}_z$, (4) $\mathbf{u}_\phi \times \mathbf{u}_x$, (5) $\mathbf{u}_\phi \times \mathbf{u}_y$, (6) $\mathbf{u}_\phi \times \mathbf{u}_z$

**QA1.26.** If $\mathbf{u}_\zeta$ is the base unit vector in the cylindrical coordinate system in the direction of the $z$ axis, what are the values of the following products? (1) $\mathbf{u}_\zeta \cdot \mathbf{u}_x$, (2) $\mathbf{u}_\zeta \cdot \mathbf{u}_y$, (3) $\mathbf{u}_\zeta \cdot \mathbf{u}_z$, (4) $\mathbf{u}_\zeta \times \mathbf{u}_x$, (5) $\mathbf{u}_\zeta \times \mathbf{u}_y$, (6) $\mathbf{u}_\zeta \times \mathbf{u}_z$

**QA1.27.** If $\mathbf{u}_x$, $\mathbf{u}_y$, and $\mathbf{u}_z$ are base unit vectors in the rectangular coordinate system and $\mathbf{u}_r$ that in the spherical coordinate system, what are the values of the following products? (1) $\mathbf{u}_r \cdot \mathbf{u}_x$, (2) $\mathbf{u}_r \cdot \mathbf{u}_y$, (3) $\mathbf{u}_r \cdot \mathbf{u}_z$, (4) $\mathbf{u}_r \times \mathbf{u}_x$, (5) $\mathbf{u}_r \times \mathbf{u}_y$, (6) $\mathbf{u}_r \times \mathbf{u}_z$

**QA1.28.** What is the physical meaning of the vector $-\nabla f$, where $f$ is a scalar function?

**QA1.29.** Vector $\mathbf{r}$ is the position vector in a scalar field described by a function $f$. What is the directional derivative of the function $f$ in the direction defined by the vector $\mathbf{r} \times \nabla f$?

**QA1.30.** What is the unit of the del operator in the cartesian (rectangular) coordinate system?

**QA1.31.** What is the divergence of the velocity field of the water flow in a pipe?

**QA1.32.** A pipe with a liquid flowing through it has a very small hole through which the liquid leaks out of the pipe. If the surface in the definition of the divergence is assumed to be finite, and if it encloses part of the pipe and the hole, is the divergence of the liquid velocity field nonzero at that point? Explain.

**QA1.33.** Propose a model in a time-variable flow of a compressible gas for which the divergence of the velocity field might be nonzero. Explain.

**QA1.34.** A fluid flows through a pipe with a rough wall. Would you expect the curl of the fluid velocity field to be nonzero? Explain.

**QA1.35.** A small spherical pressured cloud of gas is suddenly freed to disperse. Assuming completely symmetrical gas dispersion, which of the three functions of the gas velocity (a function of coordinates and time), the gradient, the divergence, and the curl, would be zero, and which nonzero? Explain.

**QA1.36.** Does the divergence theorem apply to time-dependent fields? Explain.

**QA1.37.** A volume $v$ is limited by a surface $S_0$ from outside, but has holes limited by surfaces $S_1$, $S_2$, and $S_3$ from inside. Can the divergence theorem be applied to such a domain? If it can, explain in detail the formulation of the theorem in such a case.

**QA1.38.** Is it possible to apply the divergence theorem to a vector function of the form $\mathbf{F} \times \mathbf{G}$? Explain.

**QA1.39.** Does the Stokes's theorem apply to time-dependent fields?

**QA1.40.** An open surface $S$ is limited by a large contour $C_0$, but has holes limited by small contours $C_1$ and $C_2$. Is it possible to apply Stokes's theorem to such a surface? If so, explain in detail the expression for the theorem in such a case.

**QA1.41.** Is it possible to apply Stokes's theorem to the vector function of the form $(\mathbf{F} \cdot \mathbf{G})\mathbf{F}$? Explain.

# PROBLEMS

**PA1.1.** Prove that the distributive law is valid for vector addition and for the three types of vector products (product of a vector with a scalar, dot product, and cross product).

**PA1.2.** Prove that $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A})$.

**PA1.3.** Prove that $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}$.

**PA1.4.** Let $\mathbf{n}$ be a unit vector in an arbitrary direction, and $\mathbf{r}_0$ be the position vector of a point in a plane normal to $\mathbf{n}$. What is the equation of the plane (i.e., which equation must be satisfied by the position vector of any point belonging to the plane)?

**PA1.5.** Let $\mathbf{n}$ be the unit vector along a $\zeta$ axis, and $\mathbf{r}_0$ be the position vector of the point $\zeta = 0$ on the axis. What is the expression for the position vector, $\mathbf{r}$, of a point with a coordinate $\zeta$ on the $\zeta$ axis?

**PA1.6.** Two unit vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, are in the half-plane $x > 0$ of the $z = 0$ plane. The unit vector $\mathbf{u}_1$ makes an angle $\alpha_1 > 0$ with the $x$ axis, and $\mathbf{u}_2$ an angle $\alpha_2$. Find the dot product of the two unit vectors (1) if $\alpha_1 > \alpha_2 > 0$, and (2) if $\alpha_2 < 0$. Do the results remind you of some trigonometric formulas?

**PA1.7.** Determine the unit vector in the direction of a vector $\mathbf{R}$, with the origin at a point $M'(x', y', z')$, and the tip at a point $M(x, y, z)$.

**PA1.8.** Express the base unit vectors of the cylindrical coordinate system in terms of those of the rectangular system, and conversely.

**PA1.9.** A point in the cylindrical coordinate system is defined by the coordinates $(r, \phi, z)$. Determine the rectangular coordinates of the point.

**PA1.10.** A point in the spherical coordinate system is defined by the coordinates $(r, \theta, \phi)$. Determine the rectangular coordinates of the point.

**PA1.11.** A point in the rectangular system is defined by the coordinates $(x, y, z)$. Determine the cylindrical and spherical coordinates of the point.

**PA1.12.** A vector is described at a point $M(r, \phi, z)$ in the cylindrical coordinate system by its rectangular components $A_x$, $A_y$, and $A_z$. Determine the cylindrical components of the vector.

**PA1.13.** A vector is described at a point $M(r, \theta, \phi)$ in the spherical coordinate system by its rectangular components $A_x$, $A_y$, and $A_z$. Determine the spherical components of the vector.

**PA1.14.** Given that $\mathbf{A} = A_x\mathbf{u}_x + A_y\mathbf{u}_y + A_z\mathbf{u}_z$ and $\mathbf{B} = B_x\mathbf{u}_x + B_y\mathbf{u}_y + B_z\mathbf{u}_z$, determine the smaller angle between the two vectors.

**PA1.15.** The *direction cosines* of a vector are cosines of angles between the vector and the base unit vectors. Determine the sum of the squares of the direction cosines.

**PA1.16.** If $l_A$, $m_A$, and $n_A$ are direction cosines of a vector **A**, and $l_B$, $m_B$, and $n_B$ are those of a vector **B**, find the cosine of the smaller angle between the two vectors.

**PA1.17.** A vector **A** is given by its components $A_x$, $A_y$, and $A_z$. Find the angles the vector makes with the three rectangular coordinate axes.

**PA1.18.** Express the vector product of two vectors given by their rectangular components in the form of a determinant.

**PA1.19.** Express the triple scalar product of three vectors given by their rectangular components in the form of a determinant.

**PA1.20.** Given that $\mathbf{A} = 3\mathbf{u}_x - 5\mathbf{u}_y + 8\mathbf{u}_z$, $\mathbf{B} = 4\mathbf{u}_x + \mathbf{u}_y - 3\mathbf{u}_z$, and $\mathbf{C} = -2\mathbf{u}_x + 3\mathbf{u}_y - 4\mathbf{u}_z$, determine: (1) $\mathbf{A} + \mathbf{B} - \mathbf{C}$, (2) $\mathbf{A} \cdot \mathbf{B}$, (3) $\mathbf{B} \times \mathbf{C}$, (4) $\mathbf{A} \times \mathbf{C}$, (5) $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$, (6) $(\mathbf{A} \cdot \mathbf{B})\mathbf{C}$, (7) $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}$, and (8) $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$. Find the smaller angle between **A** and **B**, and between **A** and **C**, the magnitudes of the three vectors, and the unit vectors in their direction.

**PA1.21.** If $\mathbf{A} = 2\mathbf{u}_r - \mathbf{u}_\phi + \mathbf{u}_z$, with the origin at the point $M(3, 0, 0)$ of the cylindrical coordinate system, and $\mathbf{B} = 2\mathbf{u}_r - \mathbf{u}_\phi + \mathbf{u}_z$, with the origin at the point $N(5, \pi/2, 5)$, determine: (1) $\mathbf{A} + \mathbf{B}$, (2) $\mathbf{A} \cdot \mathbf{B}$, (3) $\mathbf{A} \times \mathbf{B}$, (4) the smaller angle between the two vectors, and (5) their magnitudes.

**PA1.22.** If $\mathbf{A} = 3\mathbf{u}_r + 2\mathbf{u}_\theta - \mathbf{u}_\phi$, with the origin at the point $M(1, \pi/2, 0)$ of the spherical coordinate system, and $\mathbf{B} = -2\mathbf{u}_r - 4\mathbf{u}_\theta + 2\mathbf{u}_\phi$, with the origin at the point $N(3, \pi/2, \pi)$, determine: (1) $\mathbf{A} + \mathbf{B}$, (2) $\mathbf{A} - \mathbf{B}$, (3) $\mathbf{A} \cdot \mathbf{B}$, (4) $\mathbf{A} \times \mathbf{B}$, (5) the smaller angle between the two vectors, and (6) their magnitudes.

**PA1.23.** If $\mathbf{A} = \mathbf{u}_x + 4\mathbf{u}_y + 3\mathbf{u}_z$, and $\mathbf{B} = 2\mathbf{u}_x + \mathbf{u}_y - \mathbf{u}_z$, determine the smaller angle between the vectors, the unit vectors along the two vectors, and the ratio of their magnitudes.

**PA1.24.** Determine the differential volume and three differential areas normal to the three base unit vectors in rectangular, cylindrical, and spherical coordinate systems.

**PA1.25.** From questions QA1.9 and QA1.10, we know that the relations $\mathbf{A} \cdot \mathbf{B} = \mathbf{A} \cdot \mathbf{C}$ and $\mathbf{A} \times \mathbf{B} = \mathbf{A} \times \mathbf{C}$, taken separately, do not mean that $\mathbf{B} = \mathbf{C}$. Is this still true if *both* these relations are satisfied?

**PA1.26.** Determine the gradient of the scalar function $f(x, y, z) = x \cos 3y \exp(-4z)$, and the divergence and curl of the vector function $\mathbf{F}(x, y, z) = (2x^2 yz)\mathbf{u}_x + x \sin y \cos z\mathbf{u}_y + (x + y + z)\mathbf{u}_z$.

**PA1.27.** Prove that the curl of the gradient of the function $f(x, y, z)$ from problem PA1.26 is zero, and that the divergence of the curl of $\mathbf{F}(x, y, z)$ is zero.

**PA1.28.** Prove that the identities $\nabla \times [\nabla V(x, y, z)] = 0$ and $\nabla \cdot [\nabla \times \mathbf{A}(x, y, z)] = 0$ are satisfied for any twice differentiable functions $V(x, y, z)$ and $\mathbf{A}(x, y, z)$.

**PA1.29.** Let $\mathbf{A}(x, y, z) = xyz\mathbf{u}_x - \sin x \cos y e^z \mathbf{u}_y + xy^2 z^3 \mathbf{u}_z$. (1) Evaluate the line integral of $\mathbf{A}(x, y, z)$ around a rectangular contour in the plane $z = 0$, with a vertex at the origin, a side $a$ along the $x$ axis and a side $b$ along the $y$ axis; start from the origin along side $a$. (2) Evaluate the flux of $\nabla \times \mathbf{A}(x, y, a)$ through the contour in the direction of the base unit vector $\mathbf{u}_z$. Can you conclude that the results you obtained are correct?

**PA1.30.** Let $\mathbf{A}(x, y, z) = xyz\mathbf{u}_x + x^2 y^2 z^2 \mathbf{u}_y + x^3 y^3 z^3 \mathbf{u}_z$. (1) Evaluate the flux of $\mathbf{A}(x, y, z)$ through a cube with a vertex at the origin, and with sides of length 1 along coordinate lines for $x \geq 0$, $y \geq 0$, and $z \geq 0$. (2) Evaluate the volume integral of $\nabla \cdot \mathbf{A}(x, y, z)$ over the cube. Do you have a simple check for the accuracy of the results?

**PA1.31.** The identity in Eq. (A1.44), $\nabla \cdot (\nabla \times \mathbf{A}) = 0$, was proved by considering the closed surface $S$ having a small hole that shrinks to zero. Prove the identity by considering the closed surfaces to consist of two arbitrary open surfaces with a common boundary $C$.

**PA1.32.** The gradient of a scalar function $f$ can alternatively be defined in the form very similar to that of the divergence in Eq. (A1.22),

$$\nabla f(c_1, c_2, c_3) = \lim_{\Delta v \to 0} \frac{1}{\Delta v} \oint_{\Delta S} f \, d\mathbf{S}.$$

Using arguments similar to those for deriving the divergence in orthogonal coordinate systems, derive the analogous general expression for the gradient of $f$.

**PA1.33.** From the general formula for the gradient obtained in problem PA1.32, prove that the same expressions for the gradient in rectangular, cylindrical, and spherical coordinate systems are obtained as in section A1.4.1.

**PA1.34.** The curl of a vector function $\mathbf{F}$ can alternatively be defined in the form very similar to that of the divergence in Eq. (A1.22),

$$\nabla \times \mathbf{F}(c_1, c_2, c_3) = \lim_{\Delta v \to 0} \frac{1}{\Delta v} \oint_{\Delta S} d\mathbf{S} \times \mathbf{F}.$$

Using arguments similar to those for deriving the divergence in orthogonal coordinate systems, derive the analogous general expression for the curl of $\mathbf{F}$.

**PA1.35.** From the general formula for the curl obtained in problem PA1.34, prove that the same expressions for the curl in rectangular, cylindrical, and spherical coordinate systems are obtained as in section A1.4.3.

# Appendix 2

# Summary of Vector Identities

In the relationships described in this Appendix, **A**, **B**, **C**, **D**, and **F** are vector functions, and $V$, $W$, and $f$ are scalar functions of coordinates. It is assumed that they have all necessary derivatives.

**ALGEBRAIC IDENTITIES**

1. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
2. $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
3. $\mathbf{A} \cdot \mathbf{B} = AB \cos(\mathbf{A}, \mathbf{B})$
4. $\mathbf{A} \cdot \mathbf{A} = |\mathbf{A}|^2 = A^2$
5. $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$
6. $\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}$
7. $V(\mathbf{A} + \mathbf{B}) = V\mathbf{A} + V\mathbf{B}$
8. $\mathbf{A} \times \mathbf{B} = AB \sin(\mathbf{A}, \mathbf{B})\mathbf{n}$, where **n** is the unit vector normal to the plane of vectors **A** and **B**, and its direction is determined by the right-hand rule when vector **A** is rotated to coincide with vector **B** in the shortest way.

9. $\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$

10. $\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C}$

11. $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A})$

12. $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}$

13. $(\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = \mathbf{D} \cdot [(\mathbf{A} \times \mathbf{B}) \times \mathbf{C})] = \mathbf{C} \cdot [\mathbf{D} \times (\mathbf{A} \times \mathbf{B})] = \dots$ (using nos. 11 and 12, several other forms can be obtained)

14. $\mathbf{u}_A = \mathbf{A}/A$ (unit vector in the direction of vector $\mathbf{A}$)

## DIFFERENTIAL IDENTITIES

15. $\mathbf{u}_x \cdot \nabla V = \partial V/\partial x$ ($x$: arbitrary axis)

16. $\nabla(V + W) = \nabla V + \nabla W$

17. $\nabla(VW) = V\nabla W + W\nabla V$

18. $\nabla f(V) = f'(V)\nabla V$

19. $\nabla \cdot (\mathbf{A} + \mathbf{B}) = \nabla \cdot \mathbf{A} + \nabla \cdot \mathbf{B}$

20. $\nabla \cdot (V\mathbf{A}) = V\nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla V$

21. $\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \nabla \times \mathbf{A} - \mathbf{A} \cdot \nabla \times \mathbf{B}$

22. $\nabla \times (\mathbf{A} + \mathbf{B}) = \nabla \times \mathbf{A} + \nabla \times \mathbf{B}$

23. $\nabla \times (V\mathbf{A}) = (\nabla V) \times \mathbf{A} + V\nabla \times \mathbf{A}$

24. $\nabla \cdot (\nabla \times \mathbf{A}) = 0$

25. $\nabla \cdot (\nabla V) = \nabla^2 V = \Delta V = $ Laplacian of $V$

26. $\nabla \cdot [\nabla(VW)] = V\nabla \cdot (\nabla W) + 2\nabla V \cdot \nabla W + W\nabla \cdot (\nabla V)$

27. $\nabla \times (\nabla V) = 0$

28. $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$

## INTEGRAL IDENTITIES

### Basic integral identities

29. $\displaystyle\int_v \nabla f \, dv = \oint_S f \, d\mathbf{S}$

30. $\displaystyle\int_v \nabla \cdot \mathbf{F} \, dv = \oint_S \mathbf{F} \cdot d\mathbf{S}$    (the divergence theorem)

31. $\displaystyle\int_v \nabla \times \mathbf{F} \, dv = \oint_S d\mathbf{S} \times \mathbf{F}$

32. $\displaystyle\int_S \nabla \times \mathbf{F} \cdot d\mathbf{S} = \oint_C \mathbf{F} \cdot d\mathbf{l}$    (Stokes's theorem)

### Some integral identities derived from basic integral identities

33. $\displaystyle\oint_S \nabla \times \mathbf{A} \cdot d\mathbf{S} = 0$
(set $\mathbf{F} = \nabla \times \mathbf{A}$ in no. 30, and take into account no. 24)

**34.** $\displaystyle\int_v (V\nabla^2 W + \nabla V \cdot \nabla W)\,dv = \oint_S V\frac{\partial W}{\partial n}\,dS$

(set $\mathbf{F} = V\nabla W$ in no. 30, and use nos. 20 and 15)

**35.** $\displaystyle\int_v (V\nabla^2 W - W\nabla^2 V)\,dv = \oint_S \left(V\frac{\partial W}{\partial n} - W\frac{\partial V}{\partial n}\right)dS$

(set $\mathbf{F} = V\nabla W - W\nabla V$ in no. 30, and use nos. 20 and 15)

**36.** $\displaystyle\int_v [\nabla \times \mathbf{A} \cdot \nabla \times \mathbf{B} - \mathbf{A} \cdot \nabla \times (\nabla \times \mathbf{B})]\,dv = \oint_S (\mathbf{A} \times \nabla \times \mathbf{B}) \cdot d\mathbf{S}$

(set $\mathbf{F} = \mathbf{A} \times \nabla \times \mathbf{B}$ in no. 30, and use nos. 20 and 15)

**37.** $\displaystyle\int_v \nabla V \cdot \nabla \times \mathbf{A}\,dv = \int_S V\nabla \times \mathbf{A} \cdot d\mathbf{S}$

(set $\mathbf{F} = V\nabla \times \mathbf{A}$ in no. 30, and use nos. 20 and 24)

**38.** $\displaystyle\int_S d\mathbf{S} \times \nabla V = \oint_C V\,d\mathbf{l}$

(set $\mathbf{F} = \mathbf{C}V$ in no. 29, where $\mathbf{C}$ is a constant vector, use nos. 23 and 11, and take $\mathbf{C}$ in front of both integrals)

**39.** $\displaystyle\int_S (\nabla V \times \nabla W) \cdot d\mathbf{S} = \oint_C V\nabla W \cdot d\mathbf{l}$

(set $\mathbf{F} = V\nabla W$ in no. 29, and use nos. 23 and 27)

## GRADIENT, DIVERGENCE, CURL, AND LAPLACIAN IN ORTHOGONAL COORDINATE SYSTEMS

### Rectangular coordinate system

Notation: $f = f(x, y, z)$, $\mathbf{F} = \mathbf{F}(x, y, z)$, $F_x = F_x(x, y, z)$, $F_y = F_y(x, y, z)$, $F_z = F_z(x, y, z)$

**40.** $\displaystyle\nabla f = \frac{\partial f}{\partial x}\mathbf{u}_x + \frac{\partial f}{\partial y}\mathbf{u}_y + \frac{\partial f}{\partial z}\mathbf{u}_z$

**41.** $\displaystyle\nabla \cdot \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$

**42.** $\displaystyle\nabla \times \mathbf{F} = \mathbf{u}_x\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right) + \mathbf{u}_y\left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right) + \mathbf{u}_z\left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)$

**43.** $\displaystyle\nabla^2 f \equiv \nabla \cdot (\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$

**44.** $\displaystyle\nabla^2 \mathbf{F} = (\nabla^2 F_x)\mathbf{u}_x + (\nabla^2 F_y)\mathbf{u}_y + (\nabla^2 F_z)\mathbf{u}_z$

### Cylindrical coordinate system

Notation: $f = f(r, \phi, z)$, $\mathbf{F} = \mathbf{F}(r, \phi, z)$, $F_r = F_r(r, \phi, z)$, $F_\phi = F_\phi(r, \phi, z)$, $F_z = F_z(r, \phi, z)$

**45.** $\displaystyle\nabla f = \frac{\partial f}{\partial r}\mathbf{u}_r + \frac{1}{r}\frac{\partial f}{\partial \phi}\mathbf{u}_\phi + \frac{\partial f}{\partial z}\mathbf{u}_z$

**46.** $\nabla \cdot \mathbf{F} = \dfrac{1}{r}\dfrac{\partial(rF_r)}{\partial r} + \dfrac{1}{r}\dfrac{\partial F_\phi}{\partial \phi} + \dfrac{\partial F_z}{\partial z}$

**47.** $\nabla \times \mathbf{F} = \mathbf{u}_r\left(\dfrac{1}{r}\dfrac{\partial F_z}{\partial \phi} - \dfrac{\partial F_\phi}{\partial z}\right) + \mathbf{u}_\phi\left(\dfrac{\partial F_r}{\partial z} - \dfrac{\partial F_z}{\partial r}\right) + \mathbf{u}_z\dfrac{1}{r}\left[\dfrac{\partial(rF_\phi)}{\partial r} - \dfrac{\partial F_r}{\partial \phi}\right]$

**48.** $\nabla^2 f \equiv \nabla \cdot (\nabla f) = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial f}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2 f}{\partial \phi^2} + \dfrac{\partial^2 f}{\partial z^2}$

**49.** $\nabla^2 \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla \times (\nabla \times \mathbf{A})$

## Spherical coordinate system

Notation: $f = f(r, \theta, \phi)$, $\mathbf{F} = \mathbf{F}(r, \theta, \phi)$, $F_r = F_r(r, \theta, \phi)$, $F_\theta = F_\theta(r, \theta, \phi)$, $F_\phi = F_\phi(r, \theta, \phi)$

**50.** $\nabla f = \dfrac{\partial f}{\partial r}\mathbf{u}_r + \dfrac{1}{r}\dfrac{\partial f}{\partial \theta}\mathbf{u}_\theta + \dfrac{1}{r \sin\theta}\dfrac{\partial f}{\partial \phi}\mathbf{u}_\phi$

**51.** $\nabla \cdot \mathbf{F} = \dfrac{1}{r^2}\dfrac{\partial(r^2 F_r)}{\partial r} + \dfrac{1}{r \sin\theta}\dfrac{\partial(\sin\theta\, F_\theta)}{\partial \theta} + \dfrac{1}{r \sin\theta}\dfrac{\partial F_\phi}{\partial \phi}$

**52.** $\nabla \times \mathbf{F} = \mathbf{u}_r\dfrac{1}{r \sin\theta}\left[\dfrac{\partial(\sin\theta\, F_\phi)}{\partial \theta} - \dfrac{\partial F_\theta}{\partial \phi}\right] + \mathbf{u}_\theta\dfrac{1}{r}\left[\dfrac{1}{\sin\theta}\dfrac{\partial F_r}{\partial \phi} - \dfrac{\partial(rF_\phi)}{\partial r}\right]$

$\qquad + \mathbf{u}_\phi\dfrac{1}{r}\left[\dfrac{\partial(rF_\theta)}{\partial r} - \dfrac{\partial F_r}{\partial \theta}\right]$

**53.** $\nabla^2 f \equiv \nabla \cdot (\nabla f) = \dfrac{1}{r^2}\dfrac{\partial}{\partial r}\left(r^2\dfrac{\partial f}{\partial r}\right) + \dfrac{1}{r^2 \sin\theta}\dfrac{\partial}{\partial \theta}\left(\sin\theta\dfrac{\partial f}{\partial \theta}\right) + \dfrac{1}{r^2 \sin^2\theta}\dfrac{\partial^2 f}{\partial \phi^2}$

**54.** $\nabla^2 \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla \times (\nabla \times \mathbf{A})$

# Appendix 3

# Values of Some Important Physical Constants

| Name | Symbol | Value |
|------|--------|-------|
| Velocity of light in a vacuum | $c_0$ | $2.99792458 \cdot 10^8$ m/s |
| Absolute value of electron charge | $e$ | $(1.60210 \pm 0.00007) \cdot 10^{-19}$C |
| Magnetic moment of electron | — | $(9.2837 \pm 0.0002) \cdot 10^{-32}$A $\cdot$ m$^2$ |
| Mass of electron at rest | $m_e$ | $(9.1083 \pm 0.0003) \cdot 10^{-31}$ kg |
| Mass of neutron at rest | $m_n$ | $(1.67470 \pm 0.00004) \cdot 10^{-27}$ kg |
| Mass of proton at rest | $m_p$ | $(1.67239 \pm 0.00004) \cdot 10^{-27}$ kg |
| Permeability of a vacuum | $\mu_0$ | $4\pi \cdot 10^{-7}$H/m$(\pi = 3.14159265\ldots)$ |
| Permittivity of a vacuum | $\epsilon_0$ | $1/(\mu_0 c_0^2) = 8.85419 \cdot 10^{-12}$F/m |
| Standard acceleration of free fall | $g_g$ | $9.80665$ m/s$^2$ |
| Gravitational constant | $\gamma$ | $(6.673 \pm 0.003) \cdot 10^{-11}$N $\cdot$ m$^2$/kg$^2$ |

# Appendix 4

# Electrical Properties
# of Some Materials
# at Room Temperature
# and Low Frequencies

| Material | $\epsilon_r$ | $\sigma$(S/m) | Comment |
|---|---|---|---|
| Silver | | $6.14 \cdot 10^7$ | oxidizes |
| Copper | | $5.65 - 5.8 \cdot 10^7$ | oxidizes |
| Gold | | $4.1 \cdot 10^7$ | inert |
| Aluminum | | $3.8 \cdot 10^7$ | oxidizes |
| Tungsten | | $1.8 \cdot 10^7$ | very hard |
| Zinc | | $1.74 \cdot 10^7$ | |
| Brass | | | |
| (30% zinc) | | $1.5 \cdot 10^7$ | |
| Nickel | | $1.28 \cdot 10^7$ | |
| Bronze | | $1 \cdot 10^7$ | |

*(continued)*

| Material | $\epsilon_r$ | $\sigma$(S/m) | Comment |
|---|---|---|---|
| Iron | | $1 \cdot 10^7$ | |
| Steel | | $0.5 - 1 \cdot 10^7$ | |
| Tin | | $0.87 \cdot 10^7$ | |
| Nichrome | | $0.1 \cdot 10^7$ | |
| Graphite | | $7 \cdot 10^4$ | |
| Seawater | 70 | $3 - 5$ | |
| Wet earth | 5 to 15 | $10^{-2}$ to $10^{-3}$ | |
| Dry earth | 2 to 6 | $10^{-4}$ to $10^{-5}$ | |
| Fresh water | | | |
| (lake) | 80 | $10^{-3}$ | |
| Distilled water | 80 | $2 \cdot 10^{-4}$ | |
| Alcohol | 25 | | |
| Air | 1.006 | | breakdown $3 \, \text{kV/m}$ |
| Styrofoam | 1.03 | | |
| Teflon | 2.1 | | |
| Polystyrene | 2.56 | | |
| Rubber | 2.5 to 3 | $10^{-15}$ | |
| Paper | 2 to 4 | | |
| Quartz | 3.8 | | |
| Glass | 4 to 10 | $10^{-12}$ | |
| Mica | 5.4 | | |
| Porcelain | 6 | $10^{-10}$ | |
| Diamond | 5 to 6 | $2 \cdot 10^{-13}$ | good heat conductor |
| Silicon | 11 | | semiconductor, $\sigma$ depends on doping level |
| Gallium arsenide | 13 | | semiconductor, $\sigma$ depends on doping level |
| Barium titanate | 60 to 3600 | | anisotropic in crystalline form, dependent on mechanical conditions |

# Appendix 5

# Magnetic Properties of Some Materials

| Material | $\mu_r$ | Comment |
|---|---|---|
| Silver | 0.9999976 | diamagnetic |
| Bismuth | 0.999982 | diamagnetic |
| Copper | 0.99999 | diamagnetic |
| Gold | 0.99996 | diamagnetic |
| Water | 0.9999901 | diamagnetic |
| Aluminum | 1.000021 | paramagnetic |
| Platinum | 1.0003 | paramagnetic |
| Tungsten | 1.00008 | paramagnetic |
| Ferrite | 1000 | e.g. $NiO \cdot Fe_2O_3$, insulator |
| Ferroxcube 3 | 1500 | Mn-Zn ferrite powder |
| Cobalt | 250 | ferromagnetic |
| Nickel | 600 | ferromagnetic |
| Steel | 2000 | ferromagnetic |
| Iron (0.2 impurity) | 5000 | ferromagnetic |
| Silicon iron | 7000 | ferromagnetic |
| Purified iron (0.05 impurity) | $2 \cdot 10^5$ | ferromagnetic |
| Supermalloy | as high as $10^6$ | ferromagnetic |

# Appendix 6

# Standard (IEC) Multipliers of Fundamental Units

| Multiple | Prefix | Symbol | Example |
|----------|--------|--------|---------|
| $10^{12}$ | tera | T | $1.2\,\text{Tm} = 1.2 \cdot 10^{12}\,\text{m}$ |
| $10^{9}$ | giga | G | $12\,\text{GW} = 12 \cdot 10^{9}\,\text{W}$ |
| $10^{6}$ | mega | M | $5\,\text{MHz} = 5 \cdot 10^{6}\,\text{Hz}$ |
| $10^{3}$ | kilo | k | $22\,\text{kV} = 22 \cdot 10^{3}\,\text{V}$ |
| $10^{2}$ | hecto | h | $100\,\text{hN} = 100 \cdot 10^{2}\,\text{N}$ |
| 10 | deca | da | $32\,\text{dag} = 320\,\text{g}$ |
| $10^{-1}$ | deci | d | $2\,\text{dm} = 2 \cdot 10^{-1}\,\text{m}$ |
| $10^{-2}$ | centi | c | $75\,\text{cm} = 70 \cdot 10^{-2}\,\text{m}$ |
| $10^{-3}$ | milli | m | $56\,\text{m}\Omega = 56 \cdot 10^{-3}\,\Omega$ |
| $10^{-6}$ | micro | $\mu$ | $25\,\mu\text{H} = 25 \cdot 10^{-6}\,\text{H}$ |
| $10^{-9}$ | nano | n | $56\,\text{nA} = 56 \cdot 10^{-9}\,\text{A}$ |
| $10^{-12}$ | pico | p | $40\,\text{pF} = 40 \cdot 10^{-12}\,\text{F}$ |
| $10^{-15}$ | femto | f | $1.2\,\text{fm} = 1.2 \cdot 10^{-15}\,\text{m}$ |
| $10^{-18}$ | atto | a | $0.16\,\text{aC} = 0.16 \cdot 10^{-18}\,\text{C}$ |

# Appendix 7

# The Greek Alphabet

| Letter | | Name | Letter | | Name |
|--------|--------|---------|--------|-------------|---------|
| A | $\alpha$ | alpha | N | $\nu$ | nu |
| B | $\beta$ | beta | Ξ | $\xi$ | xi |
| Γ | $\gamma$ | gamma | O | o | omikron |
| Δ | $\delta$ | delta | Π | $\pi$ | pi |
| E | $\epsilon, \varepsilon$ | epsilon | R | $\rho, \varrho$ | rho |
| Z | $\zeta$ | zeta | Σ | $\sigma$ | sigma |
| E | $\eta$ | eta | T | $\tau$ | tau |
| Θ | $\theta, \vartheta$ | theta | Υ | $\upsilon$ | upsilon |
| I | $\iota$ | iota | Φ | $\phi, \varphi$ | phi |
| K | $\kappa$ | kappa | X | $\chi$ | chi |
| Λ | $\lambda$ | lambda | Ψ | $\psi$ | psi |
| M | $\mu$ | mu | Ω | $\omega$ | omega |

# Appendix 8

# *Theory of Lossless Metallic Waveguides*

## A8.1 General Theory of Metallic Waveguides

Although waveguides are never lossless, assuming that they have no losses greatly simplifies the analysis. We limit our analysis to cylindrical waveguides, i.e., no bends of waveguides are allowed. Consider a general lossless waveguide with a cross section as shown in Fig. 23.1 (Chapter 23), guiding a time-harmonic wave. Assume that the dielectric in the waveguide has parameters $\epsilon$ and $\mu$. Let us write Maxwell's equations in phasor (complex) form for electromagnetic waves propagating in the $+z$ direction along the waveguide.

We know that waves propagating along the waveguide, if represented in phasor (complex) form, generally have a factor $e^{-(\alpha+j\beta)z} = e^{-\gamma z}$. Because the material of the waveguide is lossless, we might be tempted to omit the attenuation coefficient, $\alpha$. However, as mentioned in the introduction to Chapter 23, below a certain frequency waveguides do not allow wave propagation, i.e., they behave as wave attenuators, so it is necessary to use the complex propagation coefficient, $\gamma$.

Because we wish to investigate the types of waves that can propagate along the $z$ axis, the sole dependence of the field vectors on $z$ must be contained in the

539

propagation factor $e^{-\gamma z}$. So the phasor field components are of the form

$$\mathbf{E}_{\text{tot}}(x, y, z) = \mathbf{E}(x, y)e^{-\gamma z} \quad \text{and} \quad \mathbf{H}_{\text{tot}}(x, y, z) = \mathbf{H}(x, y)e^{-\gamma z}, \tag{A8.1}$$

where $\mathbf{E}(x, y)$ and $\mathbf{H}(x, y)$ are complex values of the two vectors *in the plane z = 0*.

Complex Maxwell's equations in differential form contain derivatives with respect to coordinates only. The derivatives with respect to coordinates $x$ and $y$ will act on $\mathbf{E}(x, y)$ and $\mathbf{H}(x, y)$ only, and the derivative with respect to $z$ on $e^{-\gamma z}$ only. For example, for the electric field vector the derivatives with respect to $x$ and $z$ are

$$\frac{\partial \mathbf{E}_{\text{tot}}(x, y, z)}{\partial x} = \frac{\partial \mathbf{E}(x, y)}{\partial x}e^{-\gamma z} \quad \text{and} \quad \frac{\partial \mathbf{E}_{\text{tot}}(x, y, z)}{\partial z} = -\gamma \mathbf{E}(x, y)e^{-\gamma z}. \tag{A8.2}$$

Thus the factor $e^{-\gamma z}$ is common to *all* terms in Maxwell's equations, and can be canceled out, just like we canceled out the factor $e^{j\omega t}$ to obtain the phasor form of the equations. Differentiation with respect to $z$ should be replaced by $-\gamma$. Of course, once we have determined the field vectors in the plane $z = 0$, that is, $\mathbf{E}(x, y)$ and $\mathbf{H}(x, y)$, the total phasor field vectors are obtained from Eq. (A8.1).

With this in mind, let us write Maxwell's equation in differential form for $\mathbf{E} = \mathbf{E}(x, y)$ and $\mathbf{H} = \mathbf{H}(x, y)$ [we omit dependence on $(x, y)$ for brevity]. Using the expressions for curl and divergence in the rectangular coordinate system, the first Maxwell's equation in scalar form becomes

$$\frac{\partial E_z}{\partial y} - \gamma E_y = -j\omega\mu H_x, \tag{Ia}$$

$$-\gamma E_x - \frac{\partial E_z}{\partial x} = -j\omega\mu H_y, \tag{Ib}$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -j\omega\mu H_z. \tag{Ic}$$

Similarly, the second equation takes the form

$$\frac{\partial H_z}{\partial y} - \gamma H_y = j\omega\epsilon E_x, \tag{IIa}$$

$$-\gamma H_x - \frac{\partial H_z}{\partial x} = j\omega\epsilon E_y, \tag{IIb}$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = j\omega\epsilon E_z. \tag{IIc}$$

Finally, the third and fourth equations become

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} = \gamma E_z, \tag{III}$$

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} = \gamma H_z. \tag{IV}$$

If we substitute $H_y$ from Eq. (Ib) into Eq. (IIa), we obtain an equation in which $E_x$ is expressed in terms of the derivatives of $z$ components of $E_z$ and $H_z$. Similarly,

from Eqs. (I$a$) and (II$b$) we can express $E_y$ in terms of these derivatives. If next the expression for $E_y$ is substituted in Eq. (I$a$), and that for $E_x$ in Eq. (I$b$), we can also express $H_x$ and $H_y$ in terms of the derivatives of $E_z$ and $H_z$. After simple manipulations we obtain the following expressions:

$$E_x = -\frac{1}{K^2}\left(\gamma\frac{\partial E_z}{\partial x} + j\omega\mu\frac{\partial H_z}{\partial y}\right), \tag{A8.3}$$

$$E_y = -\frac{1}{K^2}\left(\gamma\frac{\partial E_z}{\partial y} - j\omega\mu\frac{\partial H_z}{\partial x}\right), \tag{A8.4}$$

$$H_x = -\frac{1}{K^2}\left(-j\omega\epsilon\frac{\partial E_z}{\partial y} + \gamma\frac{\partial H_z}{\partial x}\right), \tag{A8.5}$$

$$H_y = -\frac{1}{K^2}\left(j\omega\epsilon\frac{\partial E_z}{\partial x} + \gamma\frac{\partial H_z}{\partial y}\right), \tag{A8.6}$$

where

$$K^2 = \gamma^2 + \beta^2, \qquad \beta^2 = \omega^2\epsilon\mu. \tag{A8.7}$$

Note that the propagation coefficient $\gamma$ (and therefore also the coefficient $K$) *is not known*. So there are seven scalar unknowns (the six field components and $\gamma$).

## A8.2  Quasi-Static Nature of TEM Waves

There is an interesting general conclusion concerning TEM waves. For $H_z = 0$, Eq. (A8.3) simply states that the curl of (transversal) vector $\mathbf{E}(x, y)$ is zero. We know that this means not that $\mathbf{E}(x, y) = 0$, but that $\mathbf{E}(x, y) = -\Delta V(x, y)$, or

$$E_x(x, y) = -\frac{\partial V(x, y)}{\partial x} \quad \text{and} \quad E_y(x, y) = -\frac{\partial V(x, y)}{\partial y}. \tag{A8.8}$$

Substituting these expressions for $E_x$ and $E_y$ into Eq. (III), we obtain (FOR $E_z = 0$)

$$\frac{\partial^2 V(x, y)}{\partial x^2} + \frac{\partial^2 V(x, y)}{\partial y^2} = 0. \tag{A8.9}$$

So the electric field is derivable from a potential function which at $z = 0$ satisfies Laplace's equation in $x$ and $y$. Because boundary conditions require that the tangential $\mathbf{E}$ on conductor surfaces be zero, we reach the following conclusion: for TEM waves, *the electric field in planes where z is constant is the same as the electrostatic field corresponding to the potentials of waveguide conductors at that cross section.*

## A8.3  Derivation of General Properties of TE Wave Types

For TE waves, $E_z = 0$, so that the expressions for the transversal field components in Eqs. (A8.3) to (A8.6) reduce to

$$E_x = -\frac{j\omega\mu}{K^2}\frac{\partial H_z}{\partial y}, \tag{A8.10}$$

$$E_y = \frac{j\omega\mu}{K^2}\frac{\partial H_z}{\partial x},$$ (A8.11)

$$H_x = -\frac{\gamma}{K^2}\frac{\partial H_z}{\partial x},$$ (A8.12)

$$H_y = -\frac{\gamma}{K^2}\frac{\partial H_z}{\partial y}.$$ (A8.13)

So, we can find all the field components of a TE wave if we determine $H_z$. How can we do that? We need an equation in which $H_z$ is a single unknown. This is the Helmholtz equation (21.8). [It could, of course, also be derived from Eqs. (I–IV).] It is a vector equation representing three scalar equations in three components of the vector $\mathbf{H}$. For $\sigma = 0$ and in a rectangular coordinate system, having in mind that

$$\nabla^2[\mathbf{H}(x, y)e^{-\gamma z}] = \nabla^2[H_x(x, y)e^{-\gamma z}]\mathbf{u}_x + \nabla^2[H_y(x, y)e^{-\gamma z}]\mathbf{u}_y + \nabla^2[H_z(x, y)e^{-\gamma z}]\mathbf{u}_z,$$

the Helmholtz equation for the $H_z$ component becomes

$$\nabla^2[H_z(x, y)e^{-\gamma z}] + \omega^2\epsilon\mu H_z(x, y)e^{-\gamma z} = 0.$$

Now, the first term of the last equation is just the Laplacian of the *scalar* function $H_z e^{-\gamma z}$, and we know the Laplacian in rectangular coordinates. Omitting the dependence of $H_z$ on $(x, y)$, and having in mind Eq. (A8.7), we obtain

$$\frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + \gamma^2 H_z + \omega^2\epsilon\mu H_z = \frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + K^2 H_z = 0.$$ (A8.14)

### A8.3.1  EXPRESSIONS FOR TE WAVES IN RECTANGULAR WAVEGUIDES

The first step in determining the TE waves is to find solutions of the Helmholtz equation (A8.14), that is, $H_z(x, y)$. Once this has been determined, the field components are obtained from Eqs. (A8.10) to (A8.13).

Let us attempt to find the solution of Eq. (A8.14) in the form

$$H_z(x, y) = X(x)Y(y).$$ (A8.15)

Substituting this $H_z$ into the Helmholtz equation (A8.14), and assuming nonzero $X(x)$ and $Y(y)$, we obtain

$$\frac{1}{X(x)}\frac{d^2X(x)}{dx^2} + \frac{1}{Y(y)}\frac{d^2Y(y)}{dy^2} + K^2 = 0.$$ (A8.16)

Partial derivatives are replaced by ordinary derivatives because $X(x)$ is a function of $x$ only, and $Y(y)$ a function of $y$ only. Assume any functions $X(x)$ and $Y(y)$. After performing the indicated differentiations, the first term becomes a function of $x$ alone, for example, $f(x)$, and the second of $y$ alone, for example, $g(y)$. $f(x)$ and $g(y)$ must be such that $f(x) + g(y) + K^2 = 0$ *for all $x$ and all $y$*. How can this be achieved, having in mind that $x$ and $y$ are independent and can have *any* values? There is only one

answer: both $f(x)$ and $g(y)$ *must be equal to a constant*, for example, $f(x) = k_x^2$ and $g(y) = k_y^2$, and the sum of these constants plus $K^2$ must be zero. Thus Eq. (A8.16) can be satisfied for all $x$ and all $y$ only if

$$\frac{1}{X(x)} \frac{d^2 X(x)}{dx^2} = k_x^2, \qquad \frac{1}{Y(y)} \frac{d^2 Y(y)}{dy^2} = k_y^2,$$

or

$$\frac{d^2 X(x)}{dx^2} = k_x^2 X(x), \qquad \frac{d^2 Y(y)}{dy^2} = k_y^2 Y(y), \tag{A8.17}$$

and

$$k_x^2 + k_y^2 = K^2 = \gamma^2 + \omega^2 \epsilon \mu. \tag{A8.18}$$

Thus the propagation coefficient of TE waves is

$$\gamma = \sqrt{k_x^2 + k_y^2 - \omega^2 \epsilon \mu} = j\omega \sqrt{\epsilon \mu} \sqrt{1 - \frac{k_x^2 + k_y^2}{\omega^2 \epsilon \mu}}. \tag{A8.19}$$

The solutions of Eqs. (A8.17) are well known. They are of the form

$$X(x) = A_x \sin(k_x x) + B_x \cos(k_x x), \tag{A8.20}$$

$$Y(y) = A_y \sin(k_y y) + B_y \cos(k_y y). \tag{A8.21}$$

So the expression for $H_z(x, y)$ becomes

$$H_z(x, y) = [A_x \sin(k_x x) + B_x \cos(k_x x)][A_y \sin(k_y y) + B_y \cos(k_y y)], \tag{A8.22}$$

where $A_x, \ldots, B_y, k_x$, and $k_y$ are constants to be determined from boundary conditions.

We know that on perfectly conducting waveguide walls in Fig. 23.2, boundary conditions require that $\mathbf{E}_{\text{tang}} = 0$ and $\mathbf{H}_{\text{norm}} = 0$. From Eqs. (A8.10) to (A8.13) we see that if one of these conditions is satisfied, the other is satisfied automatically. So we request that $E_y = 0$ for $x = 0$ (the left wall in Fig. 23.2) and for $x = a$ (the right wall), and that $E_x = 0$ for $y = 0$ (the bottom wall) and for $y = b$ (the top wall). According to Eqs. (A8.10), (A8.11), and (A8.22) this means that

$$\frac{\partial H_z}{\partial x} = [A_x k_x \cos(k_x x) - B_x k_x \sin(k_x x)][A_y \sin(k_y y) + B_y \cos(k_y y)] = 0$$

$$\text{for } x = 0 \text{ and } x = a, \tag{A8.23}$$

and

$$\frac{\partial H_z}{\partial y} = [A_x \sin(k_x x) + B_x \cos(k_x x)][A_y k_y \cos(k_y y) - B_y k_y \sin(k_y y)] = 0$$

$$\text{for } y = 0 \text{ and } y = b. \tag{A8.24}$$

Eq. (A8.23) can be satisfied if $A_x = 0$ (then the cosine term in the first brackets, which is nonzero for $x = 0$, vanishes), and if

$$k_x a = m\pi, \qquad \text{or} \qquad k_x = \frac{m\pi}{a}, \qquad m = 1, 2, \ldots, \tag{A8.25}$$

since for $x = a$ and these values of $k_x$ the sine in the first brackets on the right of Eq. (A8.23) is zero.

Similarly, Eq. (A8.24) can be satisfied if $A_y = 0$, and if

$$k_y b = n\pi, \qquad \text{or} \qquad k_y = \frac{n\pi}{b}, \qquad n = 1, 2, \ldots. \tag{A8.26}$$

Noting that $\omega = 2\pi f$, the final expression for the propagation coefficient of a TE wave along a rectangular waveguide becomes

$$\gamma = j\beta, \qquad \beta = \omega\sqrt{\epsilon\mu}\sqrt{1 - \frac{f_c^2}{f^2}}, \tag{A8.27}$$

where, noting that $1/\sqrt{\epsilon\mu} = c$,

$$f_c = \frac{c}{2}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}, \qquad c = \frac{1}{\sqrt{\epsilon\mu}}. \tag{A8.28}$$

where $f_c$ is the *cutoff frequency*, as explained in Chapter 23. It depends on both the numbers $m$ and $n$, and on the waveguide dimensions.

Finally, the $H_z$ component in the cross section $z = 0$ of the waveguide is

$$H_z(x, y) = H_0 \cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{A8.29}$$

where $H_0$ is a constant depending on the level of excitation of the wave in the waveguide. The other components at $z = 0$ are obtained from Eqs. (A8.10) to (A8.13) (note that $E_z = 0$):

$$E_x(x, y) = \frac{j\omega\mu}{K^2}\frac{n\pi}{b}H_0 \cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{A8.30}$$

$$E_y(x, y) = -\frac{j\omega\mu}{K^2}\frac{m\pi}{a}H_0 \sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{A8.31}$$

$$H_x(x, y) = \frac{\gamma}{K^2}\frac{m\pi}{a}H_0 \sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{A8.32}$$

$$H_y(x, y) = \frac{\gamma}{K^2}\frac{n\pi}{b}H_0 \cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0). \tag{A8.33}$$

The values of the wave components for any $z$ are obtained by simply multiplying the preceding expressions by $e^{-\gamma z} = e^{-j\beta z}$, where $\gamma$ and $\beta$ are given in Eqs. (A8.27) and (23.21).

So we have obtained the expressions for the components of the TE wave propagating in a rectangular waveguide as functions of two integer parameters, $m = 0, 1, 2, \ldots$, and $n = 0, 1, 2, \ldots$. Evidently, not only to the field distributions for different pairs of $m$ and $n$ values differ greatly, *but also the propagation coefficient, $\beta$, is different for different pairs of numbers $m$ and $n$*. Note that $m$ represents the number of half-waves along the $x$ axis, and $n$ the number of half-waves along the $y$ axis.

We see that there is a double infinite number of TE wave types, corresponding to any possible pair of $m$ and $n$. The wave determined by a pair of numbers $m$ and $n$ is known as a *TE$_{mn}$ mode*.

# *Bibliography*

## *Introductory Level*

Attwood, S. S., *Electric and Magnetic Fields*, John Wiley & Sons, New York, 1958.

Benumof, R., *Concepts in Electricity and Magnetism*, Holt, Rinehart and Winston, New York, 1961.

Bradshaw, M. D., and W. J. Byatt, *Introductory Engineering Field Theory*, Prentice Hall, Upper Saddle River, NJ, 1967.

Carter, G. W., *The Electromagnetic Field in Its Engineering Aspects*, American Elsevier, New York, 1967.

Cheng, D. K., *Fundamentals of Engineering Electromagnetics*, Addison-Wesley, Reading, MA, 1993.

Cheston, W. B., *Elementary Theory of Electromagnetic Fields*, John Wiley & Sons, New York, 1964.

Coren, R. L., *Basic Engineering Electromagnetics: An Applied Approach*, Prentice Hall, Upper Saddle River, NJ, 1989.

Corson, D. R., and P. Lorrain, *Introduction to Electromagnetic Fields and Waves*, W. H. Freeman, San Francisco, 1962.

Coulson, C. A., *Electricity*, Oliver and Boyd, Edinburgh, 1953.

Durney, C. H., and C. C. Johnson, *Introduction to Modern Electromagnetics*, McGraw-Hill, New York, 1969.

Guru, B. S., and H. R. Hiziroglu, *Electromagnetic Field Theory Fundamentals*, PWS Publishing Company, Boston, 1998.

Harrington, R. F., *Introduction to Electromagnetic Engineering*, McGraw-Hill, New York, 1958.

Hayt, W. H., Jr., *Engineering Electromagnetics*, McGraw-Hill, New York, 1967.

Holt, C. A., *Introduction to Electromagnetic Fields and Waves*, John Wiley & Sons, New York, 1966.

Hoole, S. R. H., and P. R. P. Hoole, *A Modern Short Course in Engineering Electromagnetics*, Oxford University Press, New York, 1996.

Iskander, M. F., *Electromagnetic Fields and Waves*, Prentice Hall, Upper Saddle River, NJ, 1992.

545

Johnk, C. T. A., *Engineering Electromagnetics Fields and Waves*, John Wiley & Sons, New York, 1988, 2nd edition.

Kraus, J. D., *Electromagnetics*, McGraw-Hill, New York, 1953.

Matveev, A. N., *Electricity and Magnetism*, Mir Publishers, Moscow, 1986.

Marshall, S. V., R. E. DuBroff, and G. G. Skitek, *Electromagnetic Concepts and Applications*, Prentice Hall, Upper Saddle River, NJ, 1996, 4th edition.

Neff, H. P., Jr., *Introductory Electromagnetics*, John Wiley & Sons, New York, 1991.

Oatly, C., *Electric and Magnetic Fields; An Introduction*, Cambridge University Press, Cambridge, 1976.

Owen, G. E., *Introduction to Electromagnetic Theory*, Allyn and Bacon, Boston, 1963.

Popović, B. D., *Introductory Engineering Electromagnetics*, Addison-Wesley, Reading, MA, 1971.

Popović, B. D., *A Collection of Problems in Electromagnetism (with Solutions)*, Gradjevinska Knjiga, Belgrade, 1966 (in Serbian).

Purcell, E. M., *Electricity and Magnetism*, McGraw-Hill, New York, 1985, 2nd edition.

Rao, N. N., *Elements of Engineering Electromagnetics*, Prentice Hall, Upper Saddle River, NJ, 1977.

Reitz, J. R., and F. J. Milford, *Foundations of Electromagnetic Theory*, Addison-Wesley, Reading, MA, 1960.

Rogers, W. E., *Introduction to Electric Fields*, McGraw-Hill, New York, 1954.

Rutledge, D. B., *Electromagnetics*, Lecture Notes, Caltech, 1990.

Schelkunoff, S. A., *Electromagnetic Fields*, Blaisdell Publishing Company, New York, 1963.

Schwartz, S. E., *Electromagnetics for Engineers*, Saunders College Publishing, Philadelphia, 1990.

Shen, L. C., and J. A. Kong, *Applied Electromagnetism*, PWS Publishing Company, Boston, 1953, 3d edition.

Seely, S., *Introduction to Electromagnetic Fields*, McGraw-Hill, New York, 1958.

Skilling, H. H., *Fundamentals of Electric Waves*, John Wiley & Sons, New York, 1948.

Slater, J. C., and N. H. Frank, *Electromagnetism*, Dover Publications, New York, 1969.

Ulaby, F., *Fundamentals of Applied Electromagnetics*, Prentice Hall, Upper Saddle River, NJ, 1997.

Walsh, J. B., *Electromagnetic Theory and Engineering Applications*, The Ronald Press Company, New York, 1960.

## *Intermediate Level*

Adler, R. B., L. J. Su, and R. M. Fano, *Electromagnetic Energy Transmission and Radiation*, John Wiley & Sons, New York, 1960.

Becker, R., *Electromagnetic Fields and Interactions*, Dover Publications, New York, 1982.

Bewley, L. V., *Flux Linkages and Electromagnetic Induction*, Dover Publications, New York, 1964.

Bewley, L. V., *Two-Dimensional Fields in Electrical Engineering*, The Macmillan Company, New York, 1948.

Bohn, E. V., *Introduction to Electromagnetic Fields and Waves*, Addison-Wesley, Reading, MA, 1968.

Della Torre, E., and C. V. Longo, *The Electromagnetic Field*, Allyn and Bacon, Boston, 1969.

Fano, R. M., L. J. Chu, and R. B. Adler, *Electromagnetic Fields, Energy and Forces*, John Wiley & Sons, New York, 1960.

Hallén, E., *Electromagnetic Theory*, John Wiley & Sons, New York, 1962.

Harnwell, G. P., *Principles of Electricity and Electromagnetism*, McGraw-Hill, New York, 1949.

Javid, M., and P. M. Brown, *Field Analysis and Electromagnetics*, McGraw-Hill, New York, 1963.

King, R. W. P., *Fundamental Electromagnetic Theory*, Dover Publications, New York, 1963.

Küpfmüler, K., *Einführung in die theoretische Electrotechnik*, Springer Verlag, Berlin, 1962.

Langmuir, R. V., *Electromagnetic Fields and Waves*, McGraw-Hill, New York, 1961.

Magid, L. M., *Electromagnetic Fields, Energy, and Waves*, John Wiley & Sons, New York, 1972.

Mason, M., and W. Weaver, *The Electromagnetic Field*, Dover Publications, New York, 1929.

Maxwell, J. C., *A Treatise on Electricity and Magnetism*, Vols. 1 and 2, Dover Publications, New York, 1954.

Miner, G. F., *Lines and Electromagnetic Fields for Engineers*, Oxford University Press, New York, 1996.

Moon, P., and D. E. Spencer, *Field Theory for Engineers*, D. Van Nostrand Company, Princeton, 1961.

Nussbaum, A., *Electromagnetic Theory for Engineers and Scientists*, Prentice Hall, Upper Saddle River, NJ, 1965.

Plonsey, R., and R. E. Collin, *Principles and Applications of Electromagnetic Fields*, McGraw-Hill, New York, 1961.

Polivanov, K., *The Theory of Electromagnetic Field*, Mir Publishers, Moscow, 1983.

Ramo, S., J. R. Whinnery, and T. van Duzer, *Fields and Waves in Communication Electronics*, John Wiley & Sons, New York, 1965, 1st edition, and 1994, 3d edition.

Scott, W. T., *The Physics of Electricity and Magnetism*, John Wiley & Sons, New York, 1965.

Silvester, P., *Modern Electromagnetic Fields*, Prentice Hall, Upper Saddle River, NJ, 1968.

Slater, J. C., and N. H. Frank, *Electromagnetism*, McGraw-Hill, New York, 1947.

Surutka, J. V., *Electromagnetics*, Gradjevinska Knjiga, Belgrade, 1966 (in Serbian).

Tamm, I. E., *Fundamentals of the Theory of Electricity*, Mir Publishers, Moscow, 1979.

Weber, E., *Electromagnetic Theory: Static Fields and Their Mapping*, Dover Publications, New York, 1965.

Weiss, A. von, and H. Kleinwächter, *Übersicht über die theoretische Electrotechnik*, Geest & Portig, Leipzig, 1956.

### Advanced Level

Abraham, M., and R. Becker, *The Classical Theory of Electricity and Magnetism*, Hafner Publishing Company, New York, 1951.

Batygin, V. V., and I. N. Toptygin, *Problems in Electrodynamics*, Academic Press, New York, 1964.

Bleaney, B. I., and B. Bleaney, *Electricity and Magnetism*, Clarendon Press, Oxford, 1965.

Collin, R. E., *Field Theory of Guided Waves*, IEEE Press, New York, 1991.

DiBartolo, B., *Classical Theory of Electromagnetism*, Prentice Hall, Upper Saddle River, NJ, 1991.

Durand, E., *Electrostatique et Magnétostatique*, Masson, Paris, 1953.

Elliot, R. S., *Electromagnetics*, McGraw-Hill, New York, 1966.

Eyges, L., *The Classical Electromagnetic Field*, Dover Publications, New York, 1972.

Jackson, J. D., *Classical Electrodynamics*, John Wiley & Sons, New York, 1962.

Johnson, C. C., *Field and Wave Electrodynamics*, McGraw-Hill, New York, 1965.

Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, Prentice Hall, Upper Saddle River, NJ, 1968.

King, R. W., and Prasad, S., *Fundamental Electromagnetic Theory and Applications*, Prentice Hall, Upper Saddle River, NJ, 1986.

Kong, J. A., *Electromagnetic Wave Theory*, John Wiley & Sons, New York, 1986.

Landau, L., and E. Lifschitz, *The Classical Theory of Fields*, Addison-Wesley, Reading, MA, 1951.

O'Rahilly, A., *Electromagnetic Theory: A Critical Examination of Fundamentals*, Dover Publications, New York, 1965.

Panofsky, W. K. H., and M. Phillips, *Classical Electricity and Magnetism*, Addison-Wesley, Reading, MA, 1956.

Smythe, W. R., *Static and Dynamic Electricity*, McGraw-Hill, New York, 1968.

Sommerfeld, A., *Electrodynamics*, Academic Press, New York, 1952.

Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill, New York, 1941.

Van Bladel, J., *Electromagnetic Fields*, McGraw-Hill, New York, 1964.

Weeks, W. L., *Electromagnetic Theory for Engineering Applications*, John Wiley & Sons, New York, 1964.

# Index

548

# Units

| Quantity | Symbol | Unit | Approximate practical range |
|---|---|---|---|
| Charge | $Q$, $q^*$ | coulomb (C) | $10^{-19}$–$10$ |
| Surface charge density | $\sigma$, $\rho_s$ | coulomb per square meter (C/m$^2$) | $10^{-15}$–$10^{-5}$ |
| Voltage, potential | $V$, $v^*$ | volt (V) | $10^{-9}$–$10^8$ |
| Electric field strength | **E** | volt per meter (V/m) | $10^{-6}$–$10^7$ |
| Displacement vector | **D** | coulomb per square meter (C/m$^2$) | $10^{-15}$–$10^{-5}$ |
| Current intensity | $I$, $i^*$, ﺿ | ampere (A) | $10^{-13}$–$10^5$ |
| Current density | **J** | ampere per square meter (A/m$^2$) | $10^0$–$10^7$ |
| Capacitance | $C$ | farad (F) | $10^{-15}$–$10^0$ |
| Resistance | $R$ | ohm ($\Omega$) | $0$–$\infty$ |
| Inductance | $L$ | henry (H) | $10^{-10}$–$10^1$ |
| Conductivity | $\sigma$ | siemens per meter (S/m) | see Appendix 4 |
| Permittivity | $\epsilon$ | farad per meter (F/m) | see Appendix 4 |
| Resistivity | $\rho$ | ohm per meter ($\Omega$/m) ($\Omega m$) | see Appendix 4 |
| Permeability | $\mu$ | henry per meter (H/m) | see Appendix 5 |
| Magnetic flux density | **B** | tesla (T) | $10^{-15}$–$10^0$ |
| Magnetic field strength | **H** | ampere per meter (A/m) | $10^{-8}$–$10^7$ |
| Magnetic flux | $\Phi$ | weber (Wb) | $10^{-15}$–$10^0$ |
| Frequency | $f$ | hertz (Hz) | $0$–$10^{22}$ |
| Force | **F** | newton (N) | $10^{-9}$–$10^5$ |
| Energy, work | $W$, $A$ | joule (J) | $10^{-15}$–$10^{12}$ |
| Power (work per unit time) | $P$, $p*$ | watt (W) | $10^{-12}$–$10^{10}$ |

*In some cases it is customary to use lower-case letters for time-varying quantities.